Viewing Expert Judgment in Individual Assessments through the Lens Model: Testing the Limits of Expert Information Processing

A DISSERTATION
SUBMITTED TO THE FACULTY OF
UNIVERSITY OF MINNESOTA
BY

Martin C. Yu

IN PARTIAL FULTILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Advisor: Nathan R. Kuncel

May 2018

© Martin C. Yu 2018

Acknowledgements

When I applied to the industrial/organisational psychology graduate program at the University of Minnesota, I was particularly interested in working with my now advisor, Nathan Kuncel, for no reason other than the fact that he had a very photogenic picture on the psychology website. Looking back, I would not have done it any differently. Nathan, you have been an invaluable source of knowledge, wisdom, and encouragement, and your support and feedback has been critical in the development of the body of research that has culminated into this dissertation. I expect that this signals just the beginning of a continued and fruitful collaboration as colleagues.

I also appreciate the support and guidance from my committee: Paul Sackett, Aaron Schmidt, and John Budd. I have often felt that I have had an advisor and a half as Paul, your mentorship especially through the incredible experience of working on the College Board research projects has been instrumental in the development of my research and analytical skills. Aaron, you have always provided thoughtful feedback on my class and dissertation work, and it's great to have had someone I could chat with about Formula 1 which for some reason still hasn't caught on around here. John, I appreciate the non-I/O insight that I have gained from your class and your feedback.

I am eternally grateful to the international management consulting firm who provided the assessment data that made this dissertation possible. It has been an invaluable source of data that has allowed for an abundance of interesting analyses. I am especially appreciative of their willingness to share this data even though the results do not paint a positive picture for the firm, and I have kept their identity anonymous as a

result. For readers who are somehow able to guess the identity of this firm, please note that this data is considered legacy data that do not necessarily reflect their current assessment practices.

I have made a number of close friends during the past few years in graduate school, the closest of whom have been Ori Shewach and Tetsu Yamada. We could always be relied upon to pick each other up whenever one of us was down (figuratively and literally). As far as our cohort goes (Ori, Tetsu, Win Matsuda, Brenda Ellis, Mariah Moore), it really is as good as it gets. From our time being not at all competent in our very first class to becoming competent enough after passing comprehensive exams, we could always count on each other for assistance and commiseration. I have no doubt we all have great things ahead of us. Special thanks also to Chris Huber for convincing me to attend the program and for being a good neighbour, Amy Shu for being a buddy with which I could do judgment and decision making research, Melissa and Earl Sharpe for cleaning up after me, Allen Goebl for having a fun basement and endorsing me on Linkedin, and Geoff Lamb for being an honourary I/O and having a nice balcony.

Of course, none of this would have been possible without the support of my family: my parents, Greg and Lisa, and my sister, Melanie. The time, money, and effort that you spent on me is not something I can ever hope to be able to fully repay. You have always pushed me to achieve and to not be complacent with what I have. From an early age, I wanted to become a doctor and you have encouraged me every step along the way. I am now proud to say that it has finally happened.

Abstract

The predictive validity of any assessment system is only as good as its implementation. Across a range of decision settings, algorithmic methods of data combination often match or outperform the judgmental accuracy of expert judges. Despite this, individual assessments still largely rely on the use of expert judgment to combine candidate assessment information into an overall assessment rating to predict desired criteria such as job performance. This typically results in lower levels of validity than what could theoretically have been achieved.

Based on archival assessment data from an international management consulting firm, this dissertation presents three related studies with an overarching goal of better understanding the processes underlying why expert judgment tends to be less accurate in prediction compared to algorithmic judgmental methods. First, the Lens Model is used to break down expert judgment in individual assessments into its component processes, finding that when combining assessment information into an overall evaluation of candidates, expert assessors use suboptimal predictor weighting schemes and also use them inconsistently when evaluating multiple candidates. Second, the ability of expert assessors to tailor their judgments to maximise predictive power for specific organisations is tested by comparing models of expert judgment local and non-local to organisations. No evidence of valid expertise tailored to organisations is found as models of expert judgment local to a specific organisation performed only as well as models non-local to that organisation. Third, the importance of judgmental consistency in maximising predictive validity is evaluated by testing random weighting schemes. Here, simply

exercising *mindless consistency* by applying a randomly generated weighting scheme consistently is enough to outperform expert judgment.

Taken together, these results suggest that the suboptimal and inconsistent ways that expert assessors combine assessment information is drastically hampering their ability to make accurate evaluations of assessment candidates and to predict candidates' future job performance. Even if they are able to demonstrate valid expert insight from time to time, over the long run the opportunities for human error far outweigh any opportunity for expertise to be truly influential. Implications of these findings for how assessments are conducted in organisations as well as recommendations for how expert judgment could still be retained and improved are discussed.

Table of Contents

List of Tables

List of Figures

Overview

When combining information to predict desired outcomes, the most common approach has been to rely on clinical or expert judgment (Jeanneret & Silzer, 1998; Ryan & Sackett, 1987; Silzer & Jeanneret, 2011). Yet, mechanical methods of data combination such as using algorithms and predefined predictor weighting schemes consistently match or outperform the judgmental accuracy of clinical data combinations (i.e., subjective expert judgment) across all settings that have been studied (Grove & Meehl, 1996; Grove, Zald, Lebow, Snitz, & Nelson, 2000; Kuncel, Klieger, Connelly, & Ones, 2013). Despite mechanical methods of judgment outperforming clinical methods, people often still prefer to use and rely on expert judgment and intuition, and are resistant to using mechanical methods for reasons including overconfidence in human expertise, personal theories about how judgments should be made, and lack of knowledge about the efficacy of mechanical methods (Highhouse, 2008; Kleinmuntz, 1990; Grove & Meehl, 1996). Additionally, the general public tends to perceive clinical judgment to be more effective, and describe mechanical methods with sterile adjectives (Diab, Pui, Yankelevich, & Highhouse, 2011; Eastwood, Snook, & Luther, 2012).

In this dissertation, individual assessments are examined as a prominent example of reliance on expert judgment in combining predictor information. Across the globe, organisations make countless numbers of hiring decisions each year which, in turn, affect the lives of individuals and the success of organisations. Correctly predicting which individuals will perform to the highest level so that top performing individuals can be selected is crucial to maximising organisational productivity. Common methods for

evaluating candidates for the most complex professional and leadership positions include methods that strain the information processing limits of decision makers. Individual assessments comprise multiple assessments (e.g., simulations, role-playing, interviews, and intelligence and personality tests) that evaluate job candidates' knowledge, skills, and abilities on multiple assessment dimensions such as communication, leadership, and organisation.

Even though individual assessment information can be combined mechanically, common practice tends to rely on (and promote) the use of expert judgment (Silzer & Jeanneret, 2011). Although overall assessment ratings based on individual assessments are valid predictors of job performance (Morris, Daisley, Wheeler, & Boyer, 2015), many doubt that the information from individual assessments is being used optimally. While the importance of assessing job candidates and predicting job performance is easily agreed upon, what tends to elicit stronger opinions is with regards to how to best combine predictor information (e.g., Kuncel & Highhouse, 2011; Silzer & Jeanneret, 2011). In practice, the data combination stage of any assessment system can act to maximise predictive power or it can bottleneck the utility of the assessment system. If the reliance on expert judgment limits the predictive power of individual assessments, then improving individual assessments will require either replacing expert judgment with more mechanical methods, or improving expert judgment. Given the reluctance towards purely mechanical methods of judgment, the more realistic way forward will be to improve expert judgment, and not to outright replace expert judgment.

To lay the groundwork for improving expert judgment, the elements of expert judgment need to be broken down and key deficiencies need to be identified for improvement. In this dissertation, three studies grounded in the Lens Model (Brunswik, 1952; 1955; 1956; Hammond, 1955) are conducted in service of this goal. These studies are described below, and are also summarised in Table 1.

The first study is a complete Lens Model decomposition of individual assessments. It represents the first and most comprehensive examination of this question in the literature. The Lens Model is a classic framework for studying judgment and decision making processes that has been commonly used across many different fields. However, an examination of individual assessments using this framework has been absent from the literature. The Lens Model provides a representation of how each piece of predictor information (e.g., ratings on different assessment dimensions) is weighted by the judge(s) and combined into an overall prediction of some outcome (e.g., job performance). It additionally provides a representation of how each piece of predictor information is, in reality, related to some criterion of interest. This allows us to estimate how experts combine predictor information, and to determine how well their data combination policies match up with the actual relationships between the predictors and the criterion. The Lens Model features a number of parameters (Table 2) used to quantify various aspects of judgment, and these will be thoroughly discussed later.

The second study examines the myth of expert insight as many judges still believe that it is possible for clinical methods of data combination to outperform mechanical methods (Highhouse, 2008; Jeanneret & Silzer, 1998; Silzer & Jeanneret, 2011). One

specific application of expert insight is to tailor judgments to the characteristics that make each organisation or candidate unique. Effective use of expert insight would mean that expert assessors are able to adjust their judgmental policies in valid ways to cater to the candidates they are evaluating or to the organisation for which they are evaluating candidates. For example, they may be able to perceive unique features of candidates and collect additional information to determine how these features would affect future performance, or understand how interactions between organisational and employee characteristics can determine job performance (e.g., Silzer & Jeannert, 2011). Using the part of the Lens Model that models the judges' judgmental policies, a "model of man" (Goldberg, 1970) can be derived to estimate the weights used, on average, by the assessors to combine predictor information. Using the "model of man" approach, we can capture the judgmental policies of expert assessors at one organisation, and apply them to make judgments at another organisation. Expert judges would be expected to tailor their judgments for the specific organisation for which they conduct assessments, and this tailoring is often a selling point for assessment firms. Therefore, if expert insight is truly effective, we would expect to see that predictions made using a "model of man" at the same organisation to outperform predictions made using the same model but at a different organisation.

The third study answers the question of whether accurate predictor weighting or consistent predictor weighting is more important for maximizing predictive validity. One explanation for the superiority of mechanical methods of judgment over clinical methods is that mechanical methods can consistently apply a single set of predictor weights across

every single judgmental case. Lens Model research has shown that human judges are often inconsistent in their application of a single predictor weighting policy over multiple judgments (Karelaia & Hogarth, 2008). This third study provides a thorough examination of this issue by isolating consistent weighting from accurate weighting in making predictive judgments. Simulations are conducted to test the predictive validity of judgments made by combining individual assessment information using 1) randomly generated sets of predictor weights applied consistently across all judgmental cases, and 2) random weights applied completely randomly across all judgmental cases. These are compared against non-random judgments (expert judgment, unit weighting via simple sums, and optimal weights) to determine how effective random but consistent weights are at prediction, and how well expert judgment fares against random weighting methods.

Literature Review

Defining Mechanical and Clinical Data Combination

Mechanical Data Combination

Mechanical methods of data combination are those that rely on an objective, algorithmic method to combine information, such as equations or actuarial tables (Meehl, 1954; Sawyer, 1966; Thorndike, 1918). Here, the data combination policy is explicitly known, and the same policy is used consistently across all judgmental cases without human interference. Examples include methods of high complexity such as deriving Pareto-optimal weights to combine predictor scores in a way that achieves optimal trade-offs between desired but competing criteria (De Corte, Lievens, & Sackett, 2007), or methods of low complexity such as simply summing up all predictor scores (i.e., unit weights via simple sums).

Other terms for "mechanical" seen in the literature include: objective, actuarial, statistical, and algorithmic. These terms will be used interchangeably in this dissertation. Perhaps unsurprisingly, lay perceptions of mechanical methods commonly include a sterile and impersonal factor (e.g., rigid, inflexible, and cold; Diab, Pui, Yankelevich, & Highhouse, 2011), playing a role in the preference for judgment and decision making methods that have a more salient human component (Diab et al., 2011; Eastwood, Snook, & Luther, 2012).

Clinical Data Combination

Clinical methods of data combination are those that rely on subjective human intuition and judgment to combine information (Meehl, 1954; Sawyer, 1966; Thorndike, 1918). Here, the method of data combination is not necessarily known, as even expert judges are not always able to perfectly comprehend their own judgmental policies (Hastie & Dawes, 2001). This means that not only could data combination policies differ between judges (i.e., interrater unreliability), they could also differ between judgments within the same judge (i.e., intrarater unreliability). Simply put, the defining characteristic of a clinical data combination boils down to whether a clinician (or another type of human judge) is involved (Sawyer, 1966).

Although the term "clinical" refers to the clinical psychologists described by Meehl (1954), it is not limited to clinical psychologists. It includes anyone who engages in the use of subjective human judgment. Additionally, there is no expectation that clinical psychologists must exclusively use clinical methods of data combination; clinical information can be combined mechanically. Other terms for "clinical" seen in the literature include: subjective, holistic, intuitive, expert, and human. Again, these terms will be used interchangeably in this dissertation. The term "expert judgment" will be reserved for any discussion specific to clinical judgments made by a judge who can claim some level of expertise[1] in his or her field.

[1] Claims of expertise are often as subjective as having been judged by peers as being an expert (see the section, "The Case for Expert Judgment").

Not Just a Dichotomy

Although mechanical and clinical data combinations have often been pitted against one another, doing one does not necessarily exclude doing the other. Sawyer (1966) described mechanical synthesis, where the judge makes an overall judgment (clinically) from a given set of predictor information, after which the clinical judgment is mechanically combined with the original data to obtain a final overall judgment. He also described clinical synthesis, which is basically the opposite. In clinical synthesis, a mechanical composite is first calculated, after which the judge makes the final prediction by holistically combining the mechanical composite with the original data.

Very few studies have formally tested the effectiveness of these methods but some tentative patterns have emerged. With mechanical synthesis, including the judge's clinical prediction as one variable in the final mechanical prediction still preserves predictive accuracy relative to a purely mechanical composite that excludes the clinical prediction. On the other hand, the clinical synthesis results in poorer accuracy compared to a purely mechanical composite, but has better accuracy compared to a purely clinical composite. In both cases, combining clinical and mechanical data combination methods yields predictive validity better than clinical judgment alone, but does not outperform the purely mechanical composite.

Another approach that combines both types of judgmental methods is the bootstrapped "model of man" approach (Goldberg, 1970; Hoffman, 1960). The judge's data combination policy is modelled using regression of clinical judgment on the original predictors, after which the modelled predictor weights can then be applied back to the

original predictors to obtain a "model of man" composite judgment. This essentially allows the judge's data combination policy to be statistically estimated and then mechanically applied. Across a range of judgment and decision making scenarios, this is a method that has been found to outperform the judges themselves (i.e., their purely clinical judgment) in predicting the same criteria (Dawes, 1979; Dawes & Corrigan, 1974; Goldberg, 1970). This approach will be explored in this dissertation.

A Note on Data Combination

It should be noted that data combination is distinct from data collection. It is possible for data to be collected clinically (e.g., unstructured interview), and combined mechanically with other information using a statistical formula (Sawyer, 1966)[2]. The opposite can be true as well, where data is collected mechanically (e.g., test scores, personnel records), and combined clinically with other information by a human judge.

This dissertation is primarily concerned with the data combination stage of individual assessments. Regardless of how the component assessment information is collected, the principal goal is to compare clinical and mechanical methods of combining the same set of information into an overall judgment.

[2] Sawyer (1966) provides a thorough discussion of the interplay between data collection and data combination.

Individual Assessments

Definitions

Individual assessments (also referred to as "individual psychological assessments") are a common practice used by organisations to evaluate candidates for selection or developmental purposes, and have been considered a core competency for education and training in industrial/organizational psychology (Society for Industrial and Organizational Psychology, 2016). A smattering of ways individual assessments have been described include:

- "… one psychologist making an assessment decision for a personnel-related purpose about one individual" (Ryan & Sackett, 1987, p. 456)
- "… a process of measuring a person's knowledge, skills, abilities, and personal style to evaluate characteristics and behavior that are relevant to (predictive of) successful job performance" (Jeanneret & Silzer, 1998, p. 3)
- "… a loosely defined set of procedures, similar to practices developed and performed in clinical and counseling settings, used to make recommendations for higher level hires" (Highhouse, 2002, p. 363)
- "… assessment conducted by a highly trained individual who collects, integrates, and interprets data from a range of sources, including an interview, personality and/or cognitive testing, and business simulations. Essentially, they are individual assessment centers…" (Hazucha et al., 2011, p. 298)

- "…through the use of various tools, professionals use the individual psychological assessment to evaluate the applicant as a whole and make a prediction about the applicant's appropriateness for a position within an organization holistically" (Morris, Daisley, Wheeler, & Boyer, 2015, p. 5)
- "A defining feature of the individual assessment is the expert assessor who administers, interprets, and integrates the results of the battery (Morris et al., 2015, p. 15)

Clearly, these descriptions are quite broad. In a survey sampled from members of the Division of Industrial and Organizational Psychology of the American Psychological Association (what is now the Society for Industrial and Organizational Psychology (SIOP)), Ryan and Sackett (1987)[3] noted that across respondents, there was considerable variability in the specific assessments implemented and the assessment content dimensions for which candidates were evaluated, as well as variability in the attitudes toward the validity of these assessments. Because there is no single, specific way to implement individual assessments, it is better thought of as a class of processes or procedures for evaluating candidates for organisational selection or development purposes.

The bottom line is that candidates complete some battery of assessments (e.g., cognitive ability tests, personality tests, interviews, biodata, work sample tests, simulations), and based on their performance on these assessments, they are evaluated on

[3] Given that this study (Ryan & Sackett, 1987) is now over 30 years old, its findings may not necessarily be representative of the current state of affairs in individual assessments. An updated version of this study would be very informative and would likely garner a sizable citation count.

several job-relevant dimensions (e.g., leadership, communication, judgment). Finally, the assessor combines these dimension ratings into an overall assessment rating for each candidate to predict how well the candidate would fit the job and would perform on the job. Ideally, the assessor's judgments would be informed by a thorough job analysis of what determines effective performance for the job in question, but in reality it "…is typically as informal as a discussion with the client organization about what the job must accomplish and what competencies are required" (Thornton, Hollenbeck, & Johnson, 2010, p. 828).

Test batteries and assessment centres are two assessment methods that are similar to, but distinct from individual assessments (Silzer & Jeanneret, 2011). Test batteries consist only of psychological tests and inventories and are often administered to large numbers of candidates. Assessment centres are similar to individual assessments in that a variety of assessments are administered, candidates are evaluated on a number of job-related dimensions, and assessors as a group combine information and make an overall assessment rating. The differences are characterised by "…implementation logistics, costs, and staff requirements… [and] method differences such as the required and dominant use of simulations and multiple assessors (not necessarily psychologists) in assessment centers, whereas individual assessments often include a different, and often broader, mix of assessment tools" (Silzer & Jeanneret, 2011, p. 272). Given the overlap between these methods, research from test batteries and assessment centres have the potential to inform individual assessments and vice versa.

Validity

Overall assessment ratings from individual assessments moderately predict subjective supervisory ratings of job performance (mean corrected r = .30), and administrative decisions such as promotions, salary changes, and bonuses (mean uncorrected r = .17) (Morris et al., 2015). Although information about the validity of individual assessments at the dimension level is absent from the literature, insight into this can be drawn from research on assessment centres. Arthur, Day, McNelly, and Edens (2003) found that when predicting job performance, the validity of assessment centre dimensions fell in the range of ρ = .25 to .39, which was comparable to the predictive validity of the overall assessment rating (ρ = .36).

Given the reliance on expert judgment in combining assessment information into the overall assessment rating, it raises the question of whether they would have higher predictive validity if the assessment dimensions are instead combined into an overall rating using purely mechanical methods. In their meta-analysis, Arthur et al. (2003) found that using an optimally weighted mechanical composite of the assessment dimensions resulted in a multiple correlation of .45 with job performance, outperforming the expert-judged overall assessment rating which is typically based on a group integration meeting. It appears that although individual assessments are able to yield valid predictions of desired job performance criteria, the common practice of using the subjective judgment and intuition of expert assessors to combine assessment dimension information into an overall rating results in lower levels of predictive validity compared

to what could have theoretically been achieved with an optimally weighted mechanical data combination.

Where Did the Validity Go?

The definitions of individual assessments presented previously provide a couple of insights into the shortcomings of individual assessments and into why they demonstrate suboptimal levels of predictive validity, namely issues related to the use of human judgment. Ultimately, the levels of predictive validity that could ideally be obtained from the component assessments is let down by human error in combining assessment information into an overall evaluation of assessment candidates.

One problem has to do with who actually conducts these assessments. Ryan and Sackett (1987) defined individual assessments as being conducted by a psychologist. In practice however, this is not necessarily the case. Highhouse (2002) described the problem in very delicate terms:

> "Practitioners with no training in personnel selection and EEO issues continue to flood into this area without anyone challenging them to provide professionally acceptable evidence for the veracity of their claims. Similar to psychotherapy, it seems that the principle of functional autonomy[4] has enabled individual psychological assessment to survive and flourish" (p. 391).

[4] The term *functional autonomy* is in reference to Astin (1961), who described the increasing prevalence of psychotherapy in clinical practice without the supporting empirical evidence to back up its use.

Thornton, Hollenbeck, and Johnson (2010) informally sampled psychologists at consulting firms in 2007 and collected estimates of annual psychologist-conducted individual assessments in the United States in a range from 15,000 and 50,000[5]. Although a many of these assessments are conducted by trained psychologists, there is a lack of professional regulation in this field (Kwaske, 2004) and many assessments are conducted by non-psychologists such as managers, supervisors, trainers, and human resource professionals (Klimoski & Zukin, 2003). Drawing from assessment center research, those that used psychologists demonstrated higher predictive validity than those that used managers to evaluate candidates (Spychalski, Quiñones, Gaugler, & Pohley, 1997).

Additionally, it is not simply an issue of whether trained psychologists or non-psychologists are used as assessors. The type of training potentially matters as well. Whereas assessors trained in clinical or counseling psychology would be expected to be more skilled in interviewing and observation, assessors trained in industrial/organizational psychology would be expected to be better versed in job analysis, selection-related measurement, and legal/ethical issues in selection and assessment (Kwaske, 2004). Jeanneret and Silzer (1998) described an unpublished study at an unnamed assessment provider, where counseling psychologists were more thorough in describing their candidates, but industrial/organizational psychologists were better at predicting their candidates' future job performance.

Regardless of whether an expert or non-expert conducts these individual assessments, the fact still remains that a human judge is making the overall judgments

[5] Fairly imprecise estimates, which reflects the imprecision of expert judgment.

and predictions about candidates in individual assessments. This is evident in the definitions put forth by Morris et al. (2015), in which assessment candidates are evaluated holistically. When holistic judgment is used to combine assessment information, human error in judgment will detriment the predictive validity of individual assessments to the extent that valid use of expertise does not make up for it. The shortcomings of expert judgment using holistic methods of data combination are the focus of this dissertation, and are the subject of review in the following sections.

The Case for Expert Judgment

What Experts Do

Experts' judgmental processes have long been studied under the Naturalistic Decision Making (NDM) framework (Klein, 2008; Klein, Orasanu, Calderwood, & Zsambok, 1993). This approach attempts to decipher expert intuition and explain how expert judges make predictive judgments, often in high pressure and ill-defined situations. Here, expertise is defined by peer judgment[6] and nomination for someone who has demonstrated exceptional skill in a particular domain, often possessing some combination of experience and credentials. The use of this definition means that "[c]ompetence is inherent in the definition of expertise, so questions like "Why do experts predict badly?" do not make sense" (Philips, Klein, & Sleck, 2004, p. 299). In essence, it answers the question of what people labelled as experts do differently from those not labelled as experts, but it does not address the question of how well experts make predictive judgments as their competence is assumed.

Through the use of observation and task analysis to study expert decision making, the NDM approach resulted in the development of the recognition-primed decision model to describe the process experts use to make their judgments and decisions (Klein, Calderwood, & Clinton-Cirocco, 1986). In this model, experts 1) recognise a familiar problem within their domain of expertise, 2) search their long term memory for the first solution that intuitively comes to mind, and 3) mentally simulate the action to determine

[6] It is unclear what qualifies peers to be competent judges of one's level of expertise.

its potential efficacy. If the solution is judged to be one that works, it will be implemented, otherwise either it will be modified or another solution is generated and mentally tested. The goal here is not necessarily to obtain the optimal solution, but rather to obtain a satisfactory solution (Philips et al., 2004).

What has come out of the NDM approach to understanding expertise is the finding that experts have the requisite domain-specific knowledge to recognise familiar problems as well as the possible solutions to those problems. Therefore, expert intuition is not some mystical unknown, but rather: "The situation has provided cue: This cue has given the expert access to information stored in memory, and the information provides the answer. Intuition is nothing more and nothing less than recognition" (Simon, 1992, p. 155). To develop this expertise – or 'skilled intuition' – two conditions are required: 1) cues relevant to the problem domain must be specifiable and stable over time (i.e., the problem needs to be recognisable), and 2) there needs to be adequate opportunity to learn the cues and the responses and strategies for tackling the problem (Kahneman & Klein, 2009).

In the context of individual assessments, the implication of the recognition-primed decision model is that expert assessors should theoretically be able to recognise situations in which their expert insight can be applied to improve their evaluation of a candidate, and situations in which their judgmental policies need to be altered to suit these situations. For example, the assessor may have had prior experience with candidates who come from unconventional experiential backgrounds, and in this case evaluate a candidate with job but not educational experience differently than a candidate with

educational but not job experience. Effectively navigating through such complexities has often been touted as the hallmark of expert judgment.

Expertise in Individual Assessments

There is a heavy reliance on expert judgment in individual assessments, and Silzer and Jeanneret (2011) produced what is, to date, probably the most extensive description of all the skills and abilities that expert assessors supposedly bring to the table when conducting individual assessments. They make a number of bold claims about the use of expert judgment, including that expert assessors:

- "are accurate observers of behavior … can see and hear behavior in their observations of an individual that can provide useful and sometimes critical pieces of information to rating the individual on key dimensions" (p. 276)
- "can also formulate and test hypotheses about the individual. Using an analytical approach, they can probe and collect additional information relevant to a concern or a dimension" (p. 276)
- "can understand specific behavioral data points while also seeing larger behavioral patterns and psychological constructs" (p. 276)
- "can complete both normative and ipsative interpretations for the same variables for the same assessee that leads to a fuller understanding of that individual … a process that would be virtually impossible to complete in some mechanical or statistical manner" (p. 276)

- "can accurately sort behavior into key performance-related dimensions" (p. 277)
- "can integrate information and accurately rate an individual on specific performance dimensions" (p. 277)
- "can consider a range of behavior and determine how relevant the behavior is to later performance effectiveness" (p. 277)

In summary, Silzer and Jeanneret (2011) claim that expert assessors are able to effectively exercise their intuitive judgment to validly integrate information in complex ways. However, their assertions are not well supported by empirical evidence and "[r]esearch on decision making suggests that human judges simply cannot do the things suggested by Silzer and Jeanneret" (Kuncel & Highhouse, 2011, p. 302). This research is reviewed in the next section.

The Shortcomings of Expert Judgment

Does Expertise Actually Make a Difference?

Experts should have domain-specific knowledge and the ability and skills to implement this knowledge to solve problems. As such, they would be expected to be better at making judgments in their domain of expertise than non-experts. However, studies comparing experts and non-experts paint a less favourable picture. Goldberg (1965) examined the diagnostic accuracy of clinical psychologists and trainees relative to patient scores on the MMPI (Minnesota Multiphasic Personality Inventory), and found that both groups performed the same (r = .28). Garb (1989) reviewed studies on diagnostic accuracy in clinical practice, and concluded that while experienced clinicians and graduate students in mental health fields were better than lay judges, the clinicians were not any better than the graduate students. Therefore, it was important to have some training, but additional experience made little difference.

In more recent work, Karelaia and Hogarth's (2008) meta-analysis across a range of decision contexts noted that studies involving experts had lower validity (r = .47) than those involving trained non-experts (r = .51) or novices (r = .58). On the other hand, a meta-analysis by Spengler et al. (2009) found a small overall effect of educational and clinical experience in clinical psychology judgments (d = .12). Do note that Karelaia and Hogarth's findings were between-study comparisons, not within-study, between-group comparisons. Therefore, their set of results is purely descriptive and the groups may not

necessarily be comparable. On the other hand, Spengler et al.'s meta-analysis is based on comparing experts and non-experts performing the same decision tasks.

Similar research in the personnel assessment and selection arena is lacking, but taken together, these findings do suggest that the effects of expertise on judgmental accuracy is minimal at best. Research examining how people tend to approach judgment and decision making have uncovered a number of reasons for why the use of holistic judgment, even when implemented by experts, does not lead to more accurate judgments.

Human Judgmental Processes

Perhaps the most reliable finding in the judgment and decision making literature is that mechanical methods of data combination consistently outperform holistic methods across a variety of decision domains. This is not anything new either, as it is a finding that has dated back to as early as Meehl (1954). This includes judgments made via a range of different mechanical methods such as optimal weighting, unit weighting, and bootstrapped "model of man" weights (Grove & Meehl, 1996; Grove et al., 2000; Kuncel et al., 2013). When it comes to predicting job performance in personnel selection scenarios, mechanical methods are almost 50% more accurate than holistic methods (Kuncel et al., 2013). In their meta-analysis of Lens Model studies, Karelaia and Hogarth (2008) demonstrated that the shortcomings of expert judgment are primarily explained by inconsistency in applying predictor weights across multiple judgmental cases, and by the application of predictor weights that do not accurately reflect the predictor-criterion relationships in reality. A number of theoretical explanations for why these issues exist

have been proposed, and these are largely traced to an inappropriate use of and overconfidence in expert insight.

Process analyses of expert judgment have shown that experts tend to have some prototype in mind regarding what predictors to use when forming their judgments (Camerer & Johnson, 1991). In effect, this ends up with experts searching for and using only a subset of all the information available to match predictors to their prototype, incorporating heuristics[7] like representativeness to identify a stereotype from the predictor information, and availability regarding ease of recalling instances in which the prototype was applicable. For example, take an employer who had unsatisfactory experiences with employees who were previously self-employed or were homeschooled. Based on this prior experience, he may believe that unconventional experiential backgrounds are undesirable in the workplace, and as a result primarily search the type of employment and educational information the candidate has rather than what the candidate actually did in their previous experiences. Heuristic approaches to judgment and decision making provide shortcuts to making more efficient judgments, but the trade-off is that some predictor information may be used inconsistently, inaccurately, or even ignored from case to case.

Theoretically, reducing the number of predictor cues and using only the most relevant ones could improve expert judgment by reducing cognitive load so that judges would be able to more effectively work within the limits of working memory. Judges are

[7] A complete discussion of heuristics would be incredibly lengthy and beyond the scope of this dissertation. Interested readers are directed to Tversky and Kahneman (1974), and Shah and Oppenheimer (2008).

typically able to work with around 8 pieces of information at a time (Cooksey, 1996), and although providing more information can increase confidence in one's own judgmental accuracy, actual accuracy tends not to improve (e.g., Tsai, Klayman, & Hastie, 2008). Furthermore, providing additional information that is not relevant to the judgmental task in question has been shown to end up reducing predictive accuracy (Nisbett, Zukier, & Lemley, 1981). Meta-analytically, the average clinical judgment validity with two predictor cues is .63 and drops to .55 with three predictors, and validity is lower when there is high redundancy between predictors (r = .54) compared to no redundancy (.66; Karelaia & Hogarth, 2008). Although it seems that judgment tasks should be made as simple as possible to maximise the predictive validity of expert judgment, expert judgment tends to be anything but simple as people often attend to wrong or irrelevant predictor cues (e.g., improper consideration of broken-leg cues), or combine information in unnecessarily complex ways (e.g., attempt to account for configural rules that may not actually exist).

Broken Legs and Broken Rules

Broken-leg cues[8] (Meehl, 1954) are a potential application of expert insight in judgment and decision making processes. These are rare, but highly diagnostic cues that should theoretically override most, if not all, other predictors. Because mechanical

[8] "Broken-leg" is in reference to the example used by Meehl (1954) to describe this concept. If a man reliably goes to the movies every Thursday, the mechanical model would predict that he would see a movie next Thursday. However, if he breaks his leg on Wednesday, the human judge would be able to account for this rare occurrence and alter the prediction, now predicting that he would no longer attend the movie. Because the mechanical model does not account for this, it would result in an erroneous prediction that he would attend movie. A more common term for this concept would be "red flag."

methods typically do not account for broken-leg cues, experts who are able to detect these broken-leg cues should be expected to outperform the mechanical method. In reality however, they make little difference in prediction as by definition, they are rare and opportunities to account for them are few and far in between. Additionally, people tend to overperceive and overgeneralize the existence of these cues (Camerer & Johnson, 1991). People are drawn to them because they tell compelling stories, but this comes at a cost of potentially attending to irrelevant information, and ignoring relevant information and common sense (Highhouse, 2008). This can be considered a specific case of applying incorrect weights by overly focussing on a single piece of information (i.e., more heavily weighing one predictor over all others).

Experts often consider configural rules (i.e., interactions) between predictors to be important for maximizing their judgmental accuracy (Camerer & Johnson, 1991). This raises two questions: 1) are configural rules representative of the predictor-criterion relationships in reality, and 2) do judges actually apply configural rules effectively? The answer to both of these questions is "not really." For example, Sackett, Gruys, and Ellingson (1998) demonstrated that ability-personality interactions are rare occurrences when predicting job performance even though these reflect a common lay belief about performance and Karelaia and Hogarth (2008) showed that a linear model adequately describes how experts combine predictor values into an overall judgment, as well as

adequately describes the overall predictor-criterion relationships[9]. In a linear bootstrapped "model of man", the residual term (includes error, configural rules, and broken-leg cues) often has a negligible relationship with the criterion (Camerer & Johnson, 1991; Dawes, 1971). Yet, as is often the case with broken-leg cues, configural rules can also tell compelling stories and may also be overgeneralised (Camerer & Johnson, 1991).

The effect of using broken-leg cues and configural rules (or at least attempting to do so) results in data combination schemes that are not consistent with reality wherever and whenever these rules are irrelevant or improperly applied. Perceiving broken-leg cues or interactions where they don't actually exist will cause inappropriate and inconsistent weighting of predictors. Even in the rare cases where broken-leg cues and/or configural rules are relevant, there is no guarantee that expert judges are able to effectively account for them and to apply them only when relevant. Expert judgment is still prone to errors, and there is no evidence that experts are able to account for all of these interactions and all of the occasions in which they should be applied (Kuncel & Highhouse, 2011; Ruscio, 2003). Over the long run, with opportunities for errors being more common than opportunities for expert insight to be truly influential, the mechanical method will come out ahead.

[9] All that said, "adequate" is a subjective descriptor for some degree to which relationships are quantified by a linear model. When interactions do exist, linear models are still decent at capturing them except when they are extreme disordinal interactions. In such cases, the use of a linear model versus modelling configural rules or non-linearities is an issue of parsimony and whether fitting a more complex model results in meaningful improvements in capturing judgmental policies. This dissertation only examines linear models as there is no *a priori* expectation for specific interactions between the predictors of interest, and sample size considerations do not confidently permit a thorough search for such interactions.

Framing Effects

Framing effects provide another explanation for why expert judgment tends to be less accurate than mechanical methods of judgment. When evaluating multiple cases, mechanical methods can be made to judge each case independently, such that judgments made for one case are not influenced by judgments made for a different case. However, human judges are susceptible to framing effects that influence whether they make absolute or relative judgments.

Examples include *decoy* and *phantom* effects (Highhouse, 1997). Consider two assessment candidates (C1 and C2), with C1 scoring 8/10 and 5/10 on predictors 1 and 2 (P1 and P2), and C2 scoring 5/10 and 8/10, respectively. If the predictors are equally weighted, then the candidates are equally desirable since their overall scores are the same (13/20). Now consider that a third candidate is present and scored 8/10 on P1 and 3/10 on P2. Equally weighted, this candidate is an inferior option with an overall score of 11/20. However, because this candidate outperformed C2 on P1, but never outperformed C1 on either predictor, C1 now appears to be the better candidate relative to this *decoy* candidate.

Now consider another candidate that was present but is no longer available (e.g., accepted a different job), a *phantom* candidate, who scored 10/10 on P1 and 5/10 on P2. Equally weighted, this candidate is the superior option with an overall score of 15/20. Because the candidate is no longer available, and because people tend to be loss averse and want to minimise loss (Tversky & Kahneman, 1979; 1992), C1 also appears to be the

better candidate in this case as the loss on P1 relative to the *phantom* candidate is minimised.

The assumption in both of these examples is that the two predictors are equally desirable. A mechanical method would combine these predictor scores in exactly the same way for every candidate, and information about one candidate should not influence how the mechanical method rates another candidate. However, because human judges are susceptible to framing effects, irrelevant information is introduced that encourages relative comparisons to be made with a result of altering the judgmental policy used to combine predictor information. In essence, this can be considered a case of inappropriate use of configural rules to alter a judgmental policy in the presence of a unique informational cue. All that said, these framing effects tend to affect judgments between a small number of options, so their effects are likely localised to the options under scrutiny rather than the broader judgmental task at hand.

The Folly of Man

Unfortunately, despite the known shortcomings of expert judgment, people often still rely on clinical methods of judgment that are inconsistent and outperformed by mechanical methods. People simply prefer to make decisions and have decisions made about them based on expert judgment and intuition, and are resistant to the use of mechanical methods. A non-exhaustive list of reasons include 1) overconfidence in human expertise, 2) an assumption that complex, ill-structured problems require ill-structured methods, 3) cost of developing an appropriate algorithm, 4) availability of an

appropriate algorithm, 5) fear of being replaced by technology, 6) maintenance of self-concept and identity (i.e., it is someone's job to make decisions), 7) identifying with a specific judgmental theory, 8) viewing mechanical methods as dehumanising, and 9) poor awareness and education about the utility of mechanical methods (Highhouse, 2008; Kleinmuntz, 1990; Grove & Meehl, 1996). Among the general public, people tend to perceive clinical judgment to be more effective that mechanical methods, and they describe mechanical methods as unprofessional, impersonal, insufficient, inaccurate, unfair, and unethical (Diab et al., 2011; Eastwood et al., 2012). Clearly, pure use of mechanical methods for judgment and decision making purposes would not be a popular option.

Ultimately, if we want to improve organisational judgment and decision making processes in practice, it will require some combination of addressing these misconceptions as well as integrating expert judgment with mechanical methods rather than outright replacing expert judgment. To that end, it will be necessary to first gain a comprehensive understanding into how expert assessors use and combine information into predictive judgments. The three studies presented in this dissertation aim to break down expert judgment into its component processes and to decipher the specific processes that underlie the shortcomings of expert judgment.

Study 1: A Lens Model Decomposition of Individual Assessments

The Lens Model

In any judgmental scenario, there is some set of predictor information (i.e., cues) that relate to the criterion of interest of which the value is initially unknown. The cues provide an indication of the criterion value, and the judge estimates the criterion value by combining the information that they perceive from these cues. Essentially, the judge views the criterion through the "lens" of these predictor cues. In individual assessments, these cues are each candidate's scores on the various assessment dimensions, which are used to predict some desired criterion like job performance.

This judgmental process is captured by the Lens Model, which has long been used as a framework to analyse human judgment across a range of judgment and decision making settings. It was first conceptualised by Brunswik (1952) and then first applied towards deciphering clinical judgment by Hammond (1955). Using the Lens Model, we can evaluate the degree to which expert judgment is consistent with the actual criterion values, model how the judge weighs and combines information cues (i.e., capture their data combination policy), determine how well the judge's data combination policy matches up with the actual predictor-criterion relationships, and compare expert judgments and the judge's model with other data combination schemes. The mathematical formulation of the model that allows for this comprehensive analysis of human judgment was further developed by Hammond, Hursch, and Todd (1964), Hursch, Hammond, and Hursch (1964), and Tucker (1964).

Lens Model Parameters

The parameters of the Lens Model are shown in Figure 1, and described in detail in this section as well as in Table 2. Overall, these parameters quantify: 1) the ecology (e), which describes the criterion of interest, 2) the subject (s), which describes the clinical judgment of the criterion value, 3) the "lens", containing the independent variable or predictor cues, each bearing some relation to the criterion and providing an indication of the criterion value that informs the subject judgments, and 4) the determinants of judgmental accuracy.

Collected parameters. Three Lens Model parameters are based on collected data, and must be obtained before any other parameters can be computed. These are the predictor cues ($X_1 \ldots X_n$; where n is the number of cues), the criterion value (Y_e), and the subject response (Y_s), commonly collected as part of a validation effort. In individual assessments, the predictor cues would be each candidate's scores on the assessment dimensions, which have been distilled from their specific assessment scores. The subject response would be the expert assessor's overall assessment rating for each candidate they evaluate, and the criterion value would be some quantification of each candidate's actual job performance (e.g., supervisory ratings of performance).

The "lens". The critical function of the Lens Model is to describe how the cues are related to the criterion, and how the judge uses these cues to predict the criterion. The most common method of doing so has been to fit multiple regression models on both

sides of the lens. Often, a linear model is assumed, an assumption that has been well tested and supported (Karelaia & Hogarth, 2008).

The ecological model is a linear function describing the relationship between the cues and the criterion in the ecology, such that

$$\hat{Y}_e = b_{1e}X_1 + \cdots + b_{ne}X_n$$

where $b_{1e} \ldots b_{ne}$ are the optimal regression weights for combining the cues $X_1 \ldots X_n$ into a predicted criterion value $\hat{Y}_e$. The weights obtained from this model describe the relative value of each cue in predicting the criterion.

The subject model (i.e., bootstrapped "model of man"; see Goldberg, 1970) is a linear function describing the relationship between the cues and the subject judgment, such that

$$\hat{Y}_s = b_{1s}X_1 + \cdots + b_{ns}X_n$$

where $b_{1s} \ldots b_{ns}$ are the optimal regression weights for combining the cues $X_1 \ldots X_n$ into a predicted subject judgment $\hat{Y}_s$. The weights obtained from this model describe the relative importance placed on each predictor cue by the judge in combining the cue information into a clinical prediction of the criterion value.

Accuracy (r_a). Classically referred to as *achievement* (Hammond et al., 1964; Hursch et al., 1964), the criterion-related validity of the clinically judged criterion values in predicting the actual criterion values is simply the correlation between the two values:

$$r_a = r_{Y_e Y_s}$$

In their meta-analysis of Lens Model studies, Karelaia and Hogarth (2008) found that clinical judgment has a moderate to high degree of accuracy (average r_a = .56).

However, this is dwarfed by the validity of an optimally weighted mechanical composite (average validity = .81), meaning that whatever the judges are doing in making their clinical judgments results in predictive accuracy lower than what could have theoretically been achieved.

Tucker's (1964) alternative formulation[10] of r_a more thoroughly details the determinants of judgmental accuracy:

$$r_a = GR_eR_s + C\sqrt{(1 - R_e^2)(1 - R_s^2)}$$

Essentially, the accuracy of clinical judgment depends on the degree to which the judge combines information mechanically (GR_eR_s) plus any insight the judge utilises that is not accounted for by the mechanical model ($C\sqrt{(1 - R_e^2)(1 - R_s^2)}$). Each specific parameter in this formulation will be discussed in turn.

Environmental predictability (R_e). Perhaps evident by its name, environmental predictability quantifies the degree to which the ecological value of the criterion is predictable from a linear function of the cues. In other words, this is the criterion-related validity of an optimally weighted mechanical composite. Mathematically, this is the multiple correlation between the observed criterion values and the ecological model-predicted criterion values:

$$R_e = r_{Y_e\hat{Y}_e}$$

[10] This equation is now often referred to as *the* Lens Model equation (e.g., Karelaia & Hogarth, 2008). Ultimately, the Lens Model describes the various aspects of judgment that influence clinical judgment accuracy.

If R_e (and R_e^2, the amount of variance in the criterion accounted for by the ecological model) is high, then a linear composite of the cues yields good prediction of the criterion. If it is low, it could signal that the model is misspecified. In such a case, if the judge could account for this unmodeled knowledge, he or she could theoretically outperform the mechanical model. However, if the criterion is simply unpredictable, then the likelihood of this occurring would be expected to be nil. In reality, environmental predictability fares quite well. As mentioned before, the average R_e = .81 (Karelaia & Hogarth, 2008). Furthermore, higher R_e is correlated with higher r_a (r = .43) such that clinical judgments are more accurate if the criterion is more predictable (Karelaia & Hogarth, 2008).

Cognitive control (R_s). On the other side of the lens, cognitive control is the degree to which the subject judgments are predictable from a linear function of the cues, and is the multiple correlation between the subject judgments and the "model of man" predicted judgments:

$$R_s = r_{Y_s\hat{Y}_s}$$

Practically, what R_s (and R_s^2, the amount of variance in the subject judgments accounted for by the "model of man") captures is whether the judge applies the same linear data combination policy consistently across judgments. Karelaia and Hogarth (2008) found that cognitive control is typically quite high, where the average R_s = .80, and it has a decent correlation with r_a (r = .56). Therefore, consistently applying a single data combination policy is an important contributing factor to the accuracy of clinical judgment.

Cue sensitivity (G). This is also known as *mechanical knowledge* or a *matching index*, and is given as the correlation between the predicted criterion values and predicted judgments:

$$G = r_{\hat{Y}_e \hat{Y}_s}$$

It determines how well the "model of man" matches up with the ecological model. Because the weights in the ecology are modelled linearly, this is dependent not only on whether the judge uses weights that are consistent in magnitude and sign to the ecology, but also whether the judge applies these weights in a linear form. Cue sensitivity tends to be quite high (average G = .80; Karelaia & Hogarth, 2008), but the fact that linear models are quite robust to changes in cue weights (Dawes, 1979; Waller, 2008) suggests that it would be possible to obtain high levels of G even with weights that do not appear similar on the surface. Interestingly, G is correlated with R_s (r = .43) indicating that judges who are more consistent are better able to match the ecology, but on the other hand it is not well correlated with R_e (r = .10), indicating that judges can match the ecology regardless of its predictability.

Unmodeled knowledge (C). Although R_e and R_s are typically quite high, they are not perfect. Therefore, there is often some information that is left unmodeled in either or both the ecology and the subject. This could be random error or systematic error due to model misspecification such as: unmodeled cues, unmodeled interactions between cues, or the functional form of the model should be non-linear (Cooksey, 1996; Einhorn, 1974). An imperfect model has non-zero residual terms, which are the difference between the

observed and predicted values for each observed case. For the ecological model, the residuals are computed as:

$$Y_{e_{res}} = Y_e - \hat{Y}_e$$

For the "model of man", the residuals are computed as:

$$Y_{s_{res}} = Y_s - \hat{Y}_s$$

If the clinical judgments contain any insight not accounted for by the ecological model, this is captured as unmodeled knowledge, which is the correlation between the residuals from the two models:

$$C = r_{Y_{e_{res}} Y_{s_{res}}}$$

On average, unmodeled knowledge is very low (average C = .04), so there tends to be almost no room for improvement beyond the mechanical ecological model (Karelaia & Hogarth, 2008).

In the Lens Model equation, the unmodeled component of clinical judgment accuracy is computed as:

$$C\sqrt{(1 - R_e^2)(1 - R_s^2)}$$

This is equivalent to the covariance between the residual terms from the ecological model and from the "model of man". Functionally, this quantifies the portion of the unmodeled knowledge that contributes to predictive validity. If the environmental predictability is high, then the variance not accounted for by the ecological model $(1 - R_e^2)$ is low, so the unmodeled knowledge would make less of a difference in explaining the ecology. Similarly, if cognitive control is high, then the variance not accounted for by the "model of man" $(1 - R_s^2)$ is low, so the unmodeled knowledge would make less of a difference

in explaining the subject judgments. Additionally, if the amount of unmodeled knowledge is zero, then the Lens Model equation will simply be the modeled, mechanical component:

$$r_a = GR_eR_s \quad (\text{if } C = 0)$$

Parameter composites. From the mechanical component, a few additional and informative parameter composites can be computed.

The first is *performance* (Lindell, 1976) or *linear cognitive ability* (Hogarth & Karelaia, 2007), which "…quantifies the human, as opposed to the environmental, contribution to achievement and captures the extent to which judges both match task requirements and are consistent in the execution of their strategies" (Karelaia & Hogarth, 2008, p. 406). It is the correlation between the subject judgments and the predicted criterion values, and is equivalent to the product of the cue sensitivity and cognitive control parameters:

$$GR_s = r_{Y_s\hat{Y}_e}$$

On average, GR_s is quite high (.66), and higher than r_a (.56), so judges appear to be better at predicting the predicted criterion values than the observed criterion values (Karelaia & Hogarth, 2008).

Second is the validity of the bootstrapped "model of man" (Goldberg, 1970), which is how well the judge's average model would predict the criterion if it was applied completely consistently across all judgments. It can be computed as the product between the cue sensitivity and environmental predictability parameters:

$$GR_e = r_{Y_e\hat{Y}_s}$$

Lastly, we can subtract clinical judgmental accuracy from the validity of the "model of man":

$$GR_e - r_a$$

This provides a measure of how well judges would have done if they had applied their average data combination policy consistently instead of whatever it is they did in making their own clinical judgments. Goldberg (1970) first demonstrated the superiority of the "model of man" over the judges themselves. On average, the validity of the bootstrapped model is .65, with an average increment above clinical judgment of .10 (Karelaia & Hogarth, 2008).

Filling in the Blanks

As discussed extensively in prior sections, there is a strong reliance on clinical, expert judgment in combining information from individual assessments and concerns that this results in validity that is lower than what could be achieved if the assessment information was combined mechanically. Viewing individual assessments through the Lens Model will help to better quantify the degree to which expert judgment is detrimental to predictive validity and the underlying processes that affect the validity of expert judgment.

Additionally, the majority of Lens Model studies have been conducted in the laboratory, which likely limits the generalisability of past research. In their meta-analysis, Karelaia and Hogarth (2008) included 65 field studies out of 248 total studies (26%), so clearly there is a dearth of field studies among this literature. Furthermore, there are a

couple of indications that their results may not generalise well towards individual assessments. The predictive validity for job performance using a clinical composite of predictors is .28 (Kuncel et al., 2013), and the predictive validity of individual assessments is .30 (Morris et al., 2015). This is substantially lower than the clinical judgment accuracy found by Karelaia and Hogarth (2008), which is .56. Also, their result for the validity of an optimal mechanical composite is .81, which is almost double that of the population validity specifically for job performance using a mechanical composite of predictors at .44 (Kuncel et al., 2013).

In light of these issues, a Lens Model decomposition of individual assessments will serve to fill gaps in both individual assessment and Lens Model research. All of the aforementioned Lens Model parameters are computed using individual assessment validation datasets from two organisations, one of which conducted two separate validation studies, for a total of three validation datasets. This allows for the Lens Model to be examined between-organisations, as well as across two different samples at the same organisation.

Hypotheses

Although the magnitudes of the Lens Model parameters found in prior research may not be completely consistent with what is expected to be obtained from individual assessments, there is no prior indication that the overall patterns would not apply. Mechanical data combinations still outperform clinical combinations in predicting job performance (Kuncel et al., 2013), but the mechanical (R_e) and clinical (r_a) validities for

individual assessments would be expected to be closer to .44 and .28, respectively than .81 and .56, respectively (Hypothesis 1).

Other expectations include:

- Hypothesis 2: The bootstrapped "model of man" will outperform the experts' judgments ($GR_e - r_a > 0$)
- Hypothesis 3: Observed criterion values and subject judgments are sufficiently modelled by multiple linear regression (i.e., low unmodeled knowledge; $C < .10$)
- Hypothesis 4: Judges may be able to but are not perfect in using cue weights that are consistent with ecological weights in either or both magnitude and rank order (i.e., moderate to high cue sensitivity; $.50 < G < 1.00$)
- Hypothesis 5: Judges have some cue weighting policy that they typically use, but are inconsistent in their application of a single cue weighting policy across all judgments (i.e., moderate to high cognitive control; $.50 < R_s < 1.00$)

Method

Sample

Three archival assessment validation datasets were obtained from an international management consulting firm[11]: 1) Company A, a financial services provider (231 candidates evaluated by 26 assessors between 1994 and 1997), 2) Company B, a food retailer, Sample 1 (195 candidates evaluated by 23 assessors between 1980 and 1988), and 3) Company B, Sample 2 (421 candidates evaluated by 30 assessors between 1989 and 1999). Sample 1 and Sample 2 from Company B were obtained from separate validation studies. Candidates were evaluated for management positions by doctoral-level psychologists trained in assessment.

Based on their performance on a mix of in-basket, interviews, leaderless group discussions, personality test, and cognitive ability test, candidates were rated on seven assessment dimensions: adjustment, administration, communication, interpersonal, judgment, leadership, and motivation. Using these dimension ratings, the assessors then combined each of their candidates' ratings on these dimensions into an overall assessment rating based on person-job fit[12]. Supervisory ratings of job performance are used as the criterion variable.

[11] The identity of the consulting firm is kept anonymous as the results of this dissertation do not paint a positive picture. Their willingness to share the data that make this dissertation possible is much appreciated. For readers are somehow able to guess the identity of this firm, please note that this data is considered legacy data that do not necessarily reflect their current assessment practices.

[12] Employees who fit better with their job are expected to show higher job performance (Kristof-Brown, Zimmerman, & Johnson, 2005)

Missing data were handled by multiple imputation with predictive mean matching (Schenker & Taylor, 1996). This method randomly samples donor values from neighbouring observations that has a predicted value closest to the predicted value of the missing value. As it samples values from existing data, it maintains the plausibility of the imputed values compared to other regression-based methods. Using the MICE (Multiple Imputation by Chained Equations) package in R (van Buuren & Groothuis-Oudshoorn, 2011), five imputed datasets were generated for each of the three archival datasets, and analyses for each archival dataset were pooled across all five imputed datasets.

Analyses were conducted using both listwise deletion and multiple imputation. Conclusions were the same for both methods of handling missing data. Because listwise deletion is the less preferable option and for the sake of brevity, only results obtained via multiple imputation will be presented.

Analyses

Lens Model parameters were calculated as described in Table 2. All regression modelling were conducted using ordinary least squares multiple regression. Due to low within-assessor sample sizes, the analysis was completed at the level of each individual dataset[13].

Results were first examined for each dataset separately, and then sample-size weighted to aggregate the results across all three datasets. The Lens Model parameters

[13] This data limitation of low within-assessor sample sizes is a characteristic common to many Lens Model studies. This necessitates deriving the Lens Model for an aggregate of assessors instead of individual assessors (Karelaia & Hogarth, 2008).

obtained by Karelaia and Hogarth's (2008) meta-analysis of Lens Model studies and the clinical and mechanical validity results from Kuncel et al.'s (2013) meta-analysis were used to benchmark the parameters obtained in this study.

Results

Lens Model parameters for all three organisational validation data sources and the sample size weighted parameters are presented in Table 5, along with corresponding values from Karelaia and Hogarth (2008) and Kuncel et al. (2013).

Across all three datasets, the validity of expert judgment (accuracy; sample size weighted average r_a = .16) was far lower than what would be optimally expected from the environmental predictability (average R_e = .31). Examining the mechanical and clinical components of judgmental accuracy reveal that on average, the mechanical component contributed more to accuracy than accuracy itself (average GR_eR_s = .17), and it was the unmodeled component that was detrimental to accuracy (average $C\sqrt{(1-R_e^2)(1-R_s^2)}$ = -.01). Simply put, if the expert assessors had relied on a more mechanical approach, they would have done slightly better than they did. The exception was with Company A, where the contribution of the unmodeled component was slightly positive ($C\sqrt{(1-R_e^2)(1-R_s^2)}$ = .03), but even then, it was mostly the mechanical component that drove accuracy (GR_eR_s = .12).

The expert assessors' lack of a mechanical approach could be seen in the cognitive control parameter. Although cognitive control was relatively high (average R_s = .77), it was not perfect, so the experts' judgments were not completely linear or consistent. Cue sensitivity was also relatively high (average G = .71) but not perfect, indicating that the expert assessors have some idea of how to weigh the different assessment dimensions, but did not necessarily implement a weighting scheme close to optimal. The composites of G and R_s were modest (average GR_s = .55), so while the

assessors had somewhat defined judgmental policies, they were far from optimal and far from being applied consistently. Unmodeled knowledge was essentially zero (average C = -.02), with the negative values of unmodeled knowledge indicating that the non-linear aspects of the experts' judgments may actually be harming prediction.

Looking at the validities of the "models of man", they either matched or outperformed the expert assessors' own judgments (average GR_e = .23) in prediction. The difference was the most pronounced at Company B, Sample 1 ($GR_e - r_a$ = .16), whereas the validity of the "model of man" at Company A matched that of the assessors themselves ($GR_e - r_a$ = 0). In no case did the assessor beat their own model.

Discussion

Consistent with what has been found in the Lens Model and broader judgment and decision making literature, the Lens Model analysis of expert judgment in individual assessments conducted in this study show that on average, the predictive validity of expert judgment is far from optimal. In fact, the non-mechanical approaches to judgment used by these expert assessors to evaluate their candidates either contribute basically no additional validity or actually end up harming validity, evidenced by the near-zero or negative values for unmodeled knowledge. This study does not evaluate the reasons for why the assessors stray from a mechanical judgment approach, but rather simply finds that they do. Prior research suggests a multitude of reasons, including overconfidence in human expertise, assuming that complex problems require complex solutions, personal theories about how judgments and decisions should be made, and poor training or education (e.g., Grove & Meehl, 1996; Kleinmuntz, 1990).

That said, a subjective judgment policy is not in itself undesirable. If the experts could indeed do the things described by Silzer and Jeannert (2011) to integrate information in complex but valid ways, there may be a good deal of subjective judgment and inconsistency in judgment applied across cases as the experts evaluate information differently from case to case. The key here though, is that they do so in a manner that maximises predictive validity and not in a manner that introduces more error. Unfortunately, across all three samples studied here, this is not the case. While the experts demonstrate a decent amount of cue sensitivity, their judgmental policies are still

not close to optimal. They also demonstrate a decent amount of cognitive control, but are still far from perfect consistency.

When comparing the experts with their models (i.e., the "model of man"), they are either matched or outperformed by their model, and this exemplifies the issue of consistency. Even if the experts' judgmental policies are not optimal, simply applying their own average policies consistently across all candidates would have yielded predictive validity as good as, if not better than that of their own judgments, and would have contributed towards making a more optimal prediction. Despite the fact that such a result was observed as early as Goldberg (1970), it is unfortunate that the problem still persists decades later.

The values of the Lens Model parameters obtained in this study are found to be lower across the board compared to those from Karelaia and Hogarth's (2008) meta-analysis of Lens Model studies and Kuncel et al.'s (2013) meta-analysis of predicting job performance using clinical versus mechanical data combinations. This is especially the case for parameters directly related to predictive validity. These values were expected to be closer to Kuncel et al.'s (2013) values for clinical and mechanical judgment validity than to those from Karelaia and Hogarth (2008) as the bulk of the Lens Model meta-analysis was based on lab studies and not real-world judgment data.

Although the parameter values from this study are indeed closer to Kuncel et al. (2013), they are still a fair bit lower. This may be due to the fact that Kuncel et al.'s meta-analysis was not purely based on individual assessment judgments, but rather judgments of job performance made across a variety of predictors including cognitive ability. On

average, scores on cognitive ability measures have been shown to be the single best predictor of job performance (Schmidt & Hunter, 1998), so the inclusion of predictors with greater predictive validity than individual assessments is likely driving this difference[14]. However, the average validity of expert judgment found in this study is still lower than that of Morris et al.'s (2015) meta-analysis specific to using individual assessments to predict job performance (observed mean r = .24; corrected for criterion unreliability mean r = .30). This may simply be due to sampling error as it does fall within the 95% credibility interval, but it does suggest that the expert assessors in the three samples used for this study are performing worse than average.

By decomposing expert judgment in individual assessments into its component processes through this Lens Model analysis, it has become clear that the value of expertise in this case is essentially nil. Both the actual predictor-criterion relationships and the experts' judgments are well-captured by linear models, so there is little to no opportunity for the expert to improve upon a mechanical judgment. Even if there is room for improvement, their lack of insight into what truly constitutes an optimal judgmental policy and the inconsistency with which they evaluate their candidates detriments any inroads they may make toward beating the mechanical model. The ability of experts to validly integrate information in complex ways and the impacts of judgmental optimality and consistency are further explored in Study 2 and Study 3.

[14] Although cognitive ability tests were included in the individual assessments used for this study, their information was not directly used by the expert assessors. Rather, they were incorporated into higher-order dimension ratings. It is possible that some validity may have been lost in doing so.

Study 2: Local versus Non-Local Models of Expert Judgment

The Value and Cost of Using Expert Insight

One takeaway from Silzer and Jeanneret (2011) regarding the value of implementing individual assessments with expert judgment is that expert insight can be applied to tailor judgments to the unique characteristics of each organisation. As a result, applying expertise in such a way would be expected to maximise the predictive power of the assessment system used in each organisation. They believe that an individual assessment:

- "provides professional insight into the individual's future job performance, potential for higher-level positions, and potential to be successful in changing organizational demands" (p. 273)
- "gauges the individual's fit with the immediate manager, the peer management team, the existing direct reports and organizational structure, the organizational culture, the company values, and the country culture" (p. 273)
- "can be adapted to changing organizational contextual variables and job demands" (p. 273)

Job-specific and organisation-specific characteristics affect a variety of work-related outcomes, including job performance (Humphrey, Nahrgang, & Morgeson, 2007), and it is possible that complex relationships (e.g., non-linear relationships and configural rules) between individuals, jobs and organisations are not adequately accounted for by a simple linear weighting model. If expert assessors are able to adjust their data

combination polices to tailor their judgments of individual candidates to the job and organisational characteristics at hand in a valid manner, it would certainly be valuable to the organisation.

However, attempts to utilise expert insight would end up as being counterproductive if it introduces more error rather than predictive validity. As discussed previously, the ecology is often adequately captured by a linear model, so opportunities for expert insight to be exercised will be rare. Furthermore, in those rare cases, it is then dependent on the assessor to exercise his or her insight in a valid manner. Results from Study 1 showed that there is little unmodeled knowledge such that there is little opportunity for expert insight to add incremental prediction beyond the mechanical model. It also showed that the assessors are typically inaccurate and inconsistent when making judgments. As a result, even if the assessors attempt to use their expertise to integrate information in complex ways, over the long run, human errors take their toll on validity.

Testing the Existence of Expert Insight

Because the belief that experts provide insight into job and organizational characteristics is still a pervasive argument for the use of expert judgment in individual assessments, the aim of Study 2 is to test a specific and often touted application of expertise. This is with regard to whether expert assessors are actually able to utilise valid insight to optimally tailor their judgments for the organisations for which they conduct assessments. This involves a comparison of judgmental policies that are local and non-

local to their respective organisations. Essentially, this tests models capturing the expert assessors' judgmental policies in cross-validation. The cross-validity of optimal regression models have been examined in past research, where optimal weights applied to a new test sample still outperformed clinical judgment in prediction (Dawes, 1974; Kuncel et al., 2013).

While it has been well-established that the judge's model outperforms the judge in evaluating the same sample (Dawes, 1974; Goldberg, 197), what is unclear is whether the bootstrapped "model of man" will perform better, worse, or about the same as expert assessors in a completely new sample. This includes evaluating the predictive validity of a model derived from one sample in making predictions in a different sample, as well as evaluating the predictive validity of models derived from multiple different sources in the same sample.

In this study, average judgmental policies (i.e., "model of man") are captured using data from two organisations (organisation-specific models), and in two separate samples for one of those organisations (sample-specific models). An additional model is derived by aggregating across hundreds of organisations (general model). The local (source from which the model was derived) predictive validity and non-local validities for each model are evaluated. Additionally, the validities of these "models of man" are compared to the validities of other, more sterile (i.e., not involving any human input), mechanical methods of data combination: unit weighting via simple sums and optimal regression weights.

Hypotheses

In comparing expert judgment with the mechanical methods ("model of man", simple sums, optimal regression weights), the mechanical methods are all expected to outperform expert judgment (Hypothesis 1). Optimal regression weights are expected to perform the best, but there is an open question as to whether the "model of man" or simple sums will predict better (Hypothesis 2).

If expert insight is valid, the local model is expected to outperform any non-local model. More specifically:

- Hypothesis 3a: If expert insight is valid, the model local to one specific organisational sample should outperform the model from a different organisational sample or from a model not specific to any one organisation (i.e., the general model)

On the other hand, if expert insight is nonexistent or invalid, the local model would not be expected to outperform a non-local model, and it may be possible that by chance a non-local model ends up outperforming the local model. More specifically:

- Hypothesis 3b: If expert insight is not valid, the model local to one specific organisational sample would not outperform the model from a different organisational sample or from a model not specific to any one organisation (i.e., the general model)

Method

Sample

Study 2 uses the three archival individual assessment validation datasets from Study 1.

An additional archival, general individual assessment dataset was obtained from the same international management consulting firm, containing assessment data for 16,143 candidates evaluated by 176 assessors at 683 organisations between 1971 and 2000. It contains the dimension ratings and overall assessment ratings for each candidate, but is not a complete validation dataset as it does not contain criterion variables. Therefore, it is possible to model the subject side of the Lens Model with this dataset, but not the ecology side.

Missing data were handled using the multiple imputation procedure described in Study 1. Analyses were conducted using both listwise deletion and multiple imputation. Conclusions were the same for both methods of handling missing data. Because listwise deletion is the less preferable option and for the sake of brevity, only results obtained via multiple imputation will be presented.

Validity Analyses

An example analytical plan for Company A is depicted in Figure 2. The same analysis was conducted for Company B, Sample 1, and Company B, Sample 2, except the local and non-local assessor models will be specific to each source dataset.

Based on the method described in Study 1, a "model of man" was derived for each dataset, resulting in models capturing the average judgmental policies from four different sources: 1) Company A, 2) Company B, Sample 1, 3) Company B, Sample 2, and 3) the general assessment dataset. All four models were applied to the first three sources to combine the assessment dimensions into "model of man" overall assessment ratings, and correlated with supervisory ratings of job performance to determine each model's predictive validity for each validation source. The models were not applied to combine information from the general assessment dataset as it did not contain job performance information.

The predictive validities of overall ratings made using two other "sterile" mechanical methods – unit weighting via simple sums and optimal regression weighting – were computed to benchmark the validities of expert judgment and the "models of man" against methods that do not involve human input. Unit weighted overall ratings via simple sums were calculated by simply adding up each candidates' dimension ratings. Correlating this with their job performance yielded the predictive validity of a simple unit weighted composite. Optimal weighted overall ratings were calculated by first obtaining the optimal weights by extracting the regression coefficients from an ordinary least squares multiple linear regression model using the candidates' dimension ratings to predict their job performance. Each candidates' dimension ratings were then linearly combined using these optimal weights into an optimally weighted composite. Correlating this composite score with their job performance yielded the predictive validity of an optimally weighted composite. To adjust for sampling error in estimating the optimal

models, the correlation between optimal weighted ratings and job performance were adjusted from sample-observed values to the population level via the Wherry formula-1 (Yin & Fan, 2001).

Relative Weights Analyses

Relative weights analyses were conducted to answer two questions. First, whether it would be meaningful to tailor judgments to organisations. Answering this question involved comparing the optimal models obtained from each source to determine whether they are indeed different. The second question is whether the assessors, on average, were actually combining information differently at different organisations. This was a comparison between the "model of man" from each source to determine whether they are indeed different as the goal of this study is to determine whether observed differences in data combination policies reflect valid tailoring of judgments to specific organisations, or if it is simply due to error in judgment.

To compare models derived from different data sources, it would be necessary to evaluate the relative influence of each predictor on the whole model. The relative weight for each predictor (assessment dimension) was simply calculated by taking the proportion of weight given to each predictor relative to the sum of all weights in the model. Similarity or dissimilarity between models was determined by comparing the structure of the relative weights between models and examining whether the same dimension was weighted the same or differently in different models. This analysis was done separately for the ecology (optimal) models and the "models of man".

Results

Relative Weights

Results from the relative weights analysis for the ecology models are shown in Figure 3. Clearly, the ecology models are all different. For example, the model from Company B, Sample 2 gives an overwhelming amount of weight to the administrative dimension (relative weight = .58) whereas Company A and Company B, Sample 2 gives much less weight to the administrative dimension (relative weight = .18 and .24, respectively). Therefore, in order to maximise predictive validity at specific organisations, it would be necessary to tailor judgments towards each organisation in a valid manner.

Results from the relative weights analysis for the subject models are shown in Figure 4. Like the ecology models, it is apparent that the subject models are all different. For example, the model from Company A gives basically no weight to the motivation dimension (relative weight = .01), and Company B, Sample 2 gives little weight to the motivation dimension (relative weight = .04). The model from the general assessment dataset gives a more substantive weight to the motivation dimension (relative weight = .08), and the model from Company B, Sample 1 gives an even greater amount of weight (relative weight = .20). Therefore, the assessors are combining the assessment dimension information differently at different organisations on average. The question then is whether these differences are due to valid tailoring of judgments to organisations, or simply error.

Validity Analyses

Validities for all judgmental methods examined across all three validation datasets are presented in Table 6. At Company A (Figure 5), the predictive validity of the assessors' expert judgments (r = .17) was matched by the validity of their model (r = .17). This local model performed about the same as the non-local models (average non-local r = .17), and also about the same as unit weighting via simple sums (r = .19). Unsurprisingly, optimal weights yielded the best predictive validity (r = .28; adjusted r = .21).

At Company B, Sample 1 (Figure 6), the predictive validity of the assessors' expert judgments (r = .20) was outperformed by the validity of their model (r = .36). This local model performed about the same as the non-local models (average non-local r = .32), and also about the same as unit weighting via simple sums (r = .35). Again, optimal weights yielded the best predictive validity (r = .41; adjusted r = .37).

At Company B, Sample 2 (Figure 7), the predictive validity of the assessors' expert judgments (r = .13) was also outperformed by the validity of their model (r = .20). This local model performed about the same as the non-local models (average non-local r = .23), and also about the same as unit weighting via simple sums (r = .24). Again, optimal weights yielded the best predictive validity (r = .30; adjusted r = .28).

Discussion

The relative weights analyses demonstrated that job performance at different organisations is predicted differently by the assessment dimensions (although some of these apparent differences are likely due to sampling error). Therefore, each organisation is indeed unique with regard to job performance indicators, and assessors should tailor their judgmental policies to be in line with the dimensions most important for predicting performance. It is also the case that assessors combine information differently at different organisations as the assessor models at different organisations are also different. Ideally, these differences would be due to valid use of expertise to tailor judgments to organisations and not due to errors in judgment.

However, there does not appear to be strong evidence of valid expertise in tailoring candidate evaluations to specific organisations. Assessor models matched or outperformed the assessors themselves regardless of where the model was derived. The local models of assessors performed similar to non-local models of assessors across all organisations suggesting that expert insight, modelled as linear weights, does not appear to improve the predictive power of a mechanical combination. This was regardless of whether the local model was compared to a non-local model obtained from a different organisation, from a different sample from the same organisation, or from an aggregate of organisations.

These models of expert judgment also did not substantially outperform “sterile” mechanical methods that do not involve expert judgment (i.e., simple sums and optimal weighting). However, the advantage of using mechanical methods based on a model of

expert judgment rather than simple unit weights is that it allows us to retain expert judgment but still apply expert judgment in a way that improves predictive power over expert judgment itself. This allows experts' judgmental policies to be applied completely consistently as the experts themselves were shown to be inconsistent in their combining of dimension ratings into overall assessment ratings across candidates. That said, the source of the assessor models (i.e., local or non-local to the target organization) does not appear to be an important determinant of predictive power.

It should be noted that at Company A, the mechanical methods only marginally outperformed the assessor's expert judgments, while the mechanical methods dominated the assessors' judgments at Company B. Based on the R^2 values of the assessor models, assessors at Company A were slightly more consistent in their use of a common set of linear dimension weights across candidates than at Company B, and this better consistency may have put their judgments closer to those of mechanical methods. Other possible explanations would be that either performance at Company A was inherently more difficult to predict, or that the dimension ratings were not strong predictors of performance at Company A. Therefore, while mechanical methods of prediction were found to consistently outperform clinical methods, the degree to which they dominate may depend on the actual predictability of the criterion of interest, and the strength of the predictor-criterion relationships.

Comparing the dimension weights from the assessor models to those from the optimal models, the weights from the assessor models had little consistency with the weights for optimally predicting job performance. In other words, the assessment

dimensions considered most important by the assessors do not reflect the dimensions that are, in reality, the most important for predicting actual job performance. Given this finding, another reason why clinical methods of judgment may be less accurate than mechanical methods is that the average assessor is combining predictor information using suboptimal weights.

Based on these observations, the ability of expert assessors to produce overall candidate ratings that accurately predict candidates' job performance is impacted by both their use of suboptimal (or not-at-all optimal) weighting schemes as well as the inconsistency with which they apply their weighting schemes. The local versus non-local assessor model comparisons give the impression that the actual weights do not matter as much as the fact that mechanical methods are able to consistently apply a set of weights across all cases. Study 3 evaluates the influence of consistently weighting predictors over optimally weighting predictors in terms of maximising predictive validity.

Study 3: Comparing Random Weighting Schemes with Expert Judgments

An Inconsistent Truth

Although we have empirically come to understand experts' judgmental processes, an additional issue is that even experts tend to lack insight into their own judgmental policies (Hastie & Dawes, 2001). Clearly, it would be difficult to apply predictor weights consistently without a firm grasp of one's own judgmental policy. This is where mechanical methods of judgment shine because they are guaranteed to consistently apply a single set of predictor weights across every single judgmental case[15]. With mechanical methods, we know specifically what judgmental policy is being used and that it is being applied consistently. That said, inconsistency does not always indicate inaccuracy. For example, if the expert assessors are able to validly account for broken-leg cues, interactions, or other non-linearities, their judgmental policies will likely vary from case to case as they incorporate different pieces of information into their judgments or weigh information cues differently. The key word here, though, is "validly" as attempts to do so often end up turning out for the worse.

Studies 1 and 2 demonstrated that expert judgment is outperformed by mechanical methods of data combination, and that this is explained by both use of inaccurate weights

[15] Obviously, it would possible to create an algorithm that applies predictor weighting policies in an inconsistent manner, but as discussed previously, the operational definition of a mechanical data combination in this dissertation is the consistent application of a single data combination policy across all judgmental cases.

in combining predictor information (low cue sensitivity) as well as inconsistent use of these weighting policies (low cognitive control). Recall the Lens Model equation:

$$r_a = GR_eR_s + C\sqrt{(1 - R_e^2)(1 - R_s^2)}$$

Here, clinical judgment accuracy (r_a) is dependent on both cue sensitivity (G) and cognitive control (R_s). When judges use optimal cue weights that reflect the actual predictor-criterion relationships and/or use their weighting policy consistently, their judgmental accuracy will increase accordingly.

This raises the question of whether it is the use of accurate (optimal) weighting schemes or the consistency with which a weighting scheme is applied that drives the predictive power of a judgmental method, or if they are equally influential. Past evidence suggests that consistency is more important than optimality. As seen in Study 2, all of the mechanical models were able to match or outperform expert judgment to a similar degree in each sample regardless of the type of mechanical model or the specific predictor weights represented in each model. Linear models are robust (Dawes, 1979), meaning that changes in predictor weights do not drastically impact their predictive power as long as the signs on the weights do not change (i.e., positive weights stay positive, and negative weights stay negative). In multiple regression with three or more predictors, an infinite class of alternate regression weights (i.e., fungible weights) can be generated that yield a predictive validity approaching that of the optimal set of predictor weights (Waller, 2008). Moreover, Dawes and Corrigan (1974) found that on average, a mechanical combination using random weights applied consistently across all judgmental

cases was able to match or outperform human judges across five different judgment and decision making scenarios.

Determining the Effects of Consistency on Judgmental Accuracy

Study 3 is an extensive extension of Dawes and Corrigan (1974) to more thoroughly study the degree to which inconsistency in combining information when making multiple judgments is detrimental to the predictive validity of expert judgment. Because judgmental processes involve two aspects of data combination – the optimality of the data combination policy and the consistency with which the policy is applied – it would be necessary to tease apart consistency from optimality if the effects of consistency are to be studied. This can be done by examining random weighting schemes as there is no expectation of optimality, and pitting expert judgment against random weights in combining predictor information. When the intent is to make the most accurate judgment possible, randomly weighting information cues to make a judgment is the complete opposite of using a set of optimal regression weights.

There are two forms of random weighting that warrant consideration. The first form is the one used by Dawes and Corrigan (1974), where a set of random weights is generated, and applied consistently to every single judgmental case. In a simulation study, this is repeated many times so that the average validity of consistent use of random weights can be estimated. The second form is completely random weighting, where a set of random weights is generated for every single judgmental case. Here, no two judgments are combined using the same weighting policy (unless by coincidence). Again, this

process is repeated many times to estimate the average validity of truly random weighting. With consistent random weights, there is no expectation of optimality, but there is an expectation of consistency. With completely random weights on the other hand, there is no expectation of either optimality or consistency.

In this study, these two random weighting approaches are applied through a Monte Carlo simulation. In contrast to Dawes and Corrigan (1974), where only the average validity of consistent random weights was evaluated, this study additionally examines the average validity of completely random weights, as well as the complete distributions of validities across all simulation trials for these two random weighting approaches. By comparing the validity of subjective expert judgment and non-random mechanical methods such as unit weighting via simple sums and optimal regression weights in the context of the distributions of random but consistent and truly random weighting, we can more precisely determine the extent to which non-random methods of prediction outperform or do not outperform these random methods.

Hypotheses

Like Study 2, the optimal regression weighting method is expected to show the highest predictive validity for job performance, followed by either the "model of man" or simple sums (Hypothesis 1). All three of these methods are expected to outperform expert judgment (Hypothesis 2).

Optimal regression weighting is expected to outperform the random weighting methods in almost all cases, save any case where the random weights coincidentally

approach the optimal weights (Hypothesis 3). Unit weighting via simple sums is also expected to outperform completely random weighting in a large majority of cases (Hypothesis 4a). If sampling error in generating the consistent random weighting schemes is distributed evenly about the unit weights, unit weighting would likely be better than consistent random weighting about half the time, and worse the other half (Hypothesis 4b). Consistent application of a single set of random weights across all judgmental cases should yield more valid predictions of job performance compared to completely random weighting (Hypothesis 5).

Given the importance of consistency as discussed previously, consistent random weighting is expected to outperform expert judgment in an overwhelming majority of cases (Hypothesis 6). If we also see that completely random weights mirror the predictive power of expert judges, then we have strong evidence that the judges are using information very inconsistently and that this inconsistency in combining information does not reflect utilising expert insight and strategies specific to individuals, contexts, or jobs that improve their judgments.

Method

Sample

Study 3 uses the three archival individual assessment validation datasets from Study 1. Missing data were handled using the multiple imputation procedure described in Study 1. Analyses were conducted using both listwise deletion and multiple imputation. Conclusions were the same for both methods of handling missing data. Because listwise deletion is the less preferable option and for the sake of brevity, only results obtained via multiple imputation will be presented.

Analyses

The analyses described in this section were conducted separately using each of the three validation datasets. To simulate the use of random weights applied consistently, a set of seven weights were randomly sampled from a uniform distribution that ranged from 0 to 0.5, inclusive. This same set of random weights was then used to linearly combine each candidate's seven assessment dimension ratings into an overall assessment rating. These overall ratings were then correlated with the candidates' supervisory ratings of job performance as a measure of the predictive validity of applying a set of random weights consistently. This process is iterated 10,000 times, generating a total of 10,000 correlations as validity coefficients. Table 7 presents an example of a consistent random weighting scheme.

To simulate the use of completely random weights for each candidate, the dimension ratings for each candidate are linearly combined into an overall rating using a set of seven weights that were randomly sampled from a uniform distribution that ranges from 0 to 0.5, inclusive. A new set of seven random weights was generated to combine the dimension ratings of each candidate into overall assessment ratings. In this case, no two candidates were evaluated using the exact same weighting scheme (unless by coincidence). Again, these overall ratings were then correlated with the candidates' supervisory ratings of job performance as a measure of the predictive validity of applying completely random weights. This process is iterated 10,000 times, generating a total of 10,000 correlations as validity coefficients. An example of a completely random weighting scheme is presented in Table 7.

To provide points of comparison with non-random methods, the predictive validities of overall ratings made using non-random methods – expert judgment, simple sums, and optimal weighting – were computed. First, the overall assessment ratings made using the assessors' expert judgment were correlated with the supervisory ratings of job performance as a measure of the predictive validity of expert judgment. Second, assessor model weighted overall ratings were calculated by linearly weighting the dimension ratings by the assessor model specific to the dataset. Correlating this with the candidates' job performance yielded the predictive validity of an assessor model weighted composite. Last, optimally weighted overall ratings were calculated by first obtaining the optimal weights by extracting the regression coefficients from an ordinary least squares multiple linear regression model using the candidates' dimension ratings to predict their job

performance. Each candidates' dimension ratings were then linearly combined using these optimal weights into an optimally weighted composite. Correlating this composite score with their job performance yielded the predictive validity of an optimally weighted composite.

Results

Figure 8 displays results for analyses using the Company A data, Figure 9 displays results for analyses using the Company B, Sample 1 data, and Figure 10 displays results for analyses using the Company B, Sample 2 data.

Comparing the non-random methods in predicting supervisory ratings of job performance at Company A, overall ratings made using optimal weights (r = .25) were better predictors than those made using simple sums (r = .19), which in turn performed about the same as those made using clinical expert judgment (r = .17). In Company B, Sample 1, optimal weights (r = .40) were better than unit weights (r = .33), which were better than expert judgment (r = .16), and a similar pattern was found in Company B, Sample 2 where optimal weights (r = .30) were better than unit weights (r=.22), which were better than expert judgment (r= .13).

When the overall ratings computed using random methods were used to predict job performance at Company A, across 10,000 iterations, random weights applied consistently across candidates had a mean predictive validity of r = .18 (SD = .02), and ranged from r = .10 to .22. Random weights applied consistently outperformed expert judgments in 76.83% of the iterations, simple sums in 39.40% of the iterations, and never outperformed optimal weights. Completely random weighting across candidates had a mean validity of r = .09 (SD = .02), and ranged from r = -.01 to .19. Completely random weights never outperformed expert judgment, simple sums, or optimal weights. 69.85% of the iterations for completely random weights were outperformed by all of the iterations for random weights applied consistently.

At Company B, Sample 1 across 10,000 iterations, random weights applied consistently across candidates had a mean validity of r = .34 (SD = .03), and ranged from r = .20 to .40. Random weights applied consistently outperformed expert judgments in 100% of the iterations, simple sums in 32.96% of the iterations, and never outperformed optimal weights. Completely random weighting across candidates had a mean validity of r = .16 (SD = .03), and ranged from r = .05 to .27. Completely random weights outperformed expert judgments in 8.49% of the iterations, but never outperformed simple sums or optimal weights. 94.05% of the iterations for completely random weights were outperformed by all of the iterations for random weights applied consistently.

At Company B, Sample 2 across 10,000 iterations, random weights applied consistently across candidates had a mean validity of r = .24 (SD = .02), and ranged from r = .15 to .29. Random weights applied consistently outperformed expert judgments in 100% of the iterations, simple sums in 36.12% of the iterations, and never outperformed optimal weights. Completely random weighting across candidates had a mean validity of r = .12 (SD = .02), and ranged from r = .05 to .20. Completely random weights outperformed expert judgments in 22.22% of the iterations, but never outperformed simple sums or optimal weights. 96.69% of the iterations for completely random weights were outperformed by all of the iterations for random weights applied consistently.

Discussion

Across the three samples, experts outperformed completely random weights 100%, 91.5%, and 77.8% of the time in predicting subsequent job performance ratings in Company A, Company B, Sample 1, and Company B, Sample 2, respectively. In turn, random weights, consistently applied, outperformed experts 76.8%, 100%, and 100% of the time. These results indicate that experts do not make judgments completely randomly and are aware, to some extent, of what information is most valuable. However, their inconsistency in combining information does drastically damage their accuracy. This simulation study demonstrates that consistency in applying predictor weights is paramount to making accurate judgments.

There are several points that are worth considering. The most important and striking is that *mindless consistency* is enough to result in more accuracy than expert judgment. On average, random weights applied consistently resulted in better predictions than the assessors' own judgments, which parallels Dawes and Corrigan's (1974) earlier study of random weighting. In the Company A analyses, consistent use of random weights dominated the experts in the majority of cases. In the analyses for both Samples 1 and 2 at Company B, consistent use of random weights completely dominated the experts.

At this point, it is unclear what determines the extent to which consistent random weights will dominate over expert judgment, but it may be in part a function of the strength of the predictor-criterion relationships: based on the predictive validity of the optimal weighting schemes, it is clear that these assessment dimensions are better at

predicting performance at Company B than at Company A. However, both Company B samples show 100% dominance of consistent random weights over expert judgment, but the optimal validity and the validity of expert judgment at Sample 2 are both lower than those at Sample 1. Another possibility is that dominance is dependent on the difference between the validities of expert judgment and optimal weighting as this difference is larger in both Company B samples compared to Company A. More research across a larger number of samples will be needed to decipher the mechanism underlying this dominance effect.

Differences in the validity of expert judgment across the three samples were fairly small. However, the dominance of expert judgment over completely random weighting was not the same. At Company A, expert judgment was completely better than completely random weights, but at both Company B samples, expert judgment was not always better than completely random. Possible explanations include that the assessors at Company A were simply more consistent or incorporating mechanical approaches to their judgment, or there may be differences in the variability of candidate characteristics between each sample that may impact how well a completely random weighting scheme would perform. Further research will be needed to determine organizational and individual differences that may influence the differences in validity between clinical expert judgment and mechanical methods of judgment. Nevertheless, it is troubling that expert assessors are not always better than completely random as it suggests that they do not necessarily understand what they are doing when combining information and evaluating candidates.

Ultimately, the finding that even random weights perform well when applied consistently suggests that consistency in applying predictor weights are more important than the weights themselves. Linear models are quite robust, and as long as the signs on the weights do not change (as is the case in the present study where all weights were positive), changes in weights are not expected to drastically impact their predictive power (Dawes, 1979). As Waller (2008) demonstrated with fungible weights, it is possible to derive an infinite number of alternate regression weighting schemes that yield a predictive validity almost as good as that of optimal weights (in multiple regression with three or more predictors). That being said, even though it is possible to generate a set of random weights that will perform very well when applied consistently, it can be difficult or impossible to tell how well that set of random weights will perform until the validation is conducted. In this simulation study, both optimal and unit weights via simple sums tend to perform better than random weights applied consistently. Practically speaking, if optimal weights are not known or cannot be approximated, it would be better to simply add up predictor scores instead of using an ill-defined weighting scheme.

General Discussion

Organisations that implement individual assessments have largely relied on expert judgment to evaluate the suitability of candidates for jobs and to predict candidates' future job performance. This is a practice that has been advocated by prominent members of our field such a Silzer and Jeanneret (2011), who made weakly supported claims that expert judges can integrate information in superhuman ways. What has been demonstrated in this study is that in comparison to the levels of predictive validity that could theoretically have been achieved, the expert assessors in fact typically do quite poorly when evaluating assessment candidates.

Study 1 was a Lens Model decomposition of expert judgment in individual assessments. It found that the validity of expert judgment was, on average, far from optimal and that experts used suboptimal judgment policies and used them inconsistently. If the experts had basically used their own policy consistently, they would have done better than whatever it was that they did when making their original judgments. Study 2 examined whether the expert assessors were able to validly tailor their judgmental policies to specific organisations to maximise prediction at specific organisations. The opposite was found where judgment policies local and non-local to organisations all performed about the same no matter which organisation they were utilised. Additionally, simply adding up predictor scores performed just as well. Study 3 determined that judgmental consistency was paramount to maximising predictive validity by simulating random weighting schemes. Even *mindless consistency* by applying a random weighting

policy consistently across all candidates was enough to outperform the experts. Taken together, the three studies presented in this dissertation along with the broader judgment and decision making literature demonstrate that claims about the ability of expert judges to handle complex information and to make complex judgments are not grounded in reality. Over the long run, even if the expert can beat the mechanical model from time to time, opportunities for human error will be a lot more common than opportunities for expertise to be truly influential. As a result, the mechanical model will come out ahead.

Making an Expert

It could be argued that the assessors present in the samples studied are not truly experts, and this boils down to an issue of how expertise is determined. Because the judgmental accuracy of the assessors cannot be determined until after the validation study has been conducted, it is the assumption of expertise that is of critical concern, which is an assumption based on credentials, experience, and/or nomination by peers. This definition of expertise is similar to that used in the Naturalistic Decision Making paradigm for studying expert judgment (Klein, 2008; Klein et al., 1993). With this approach, expertise is assumed and it is assumed that the judgments and decisions of these assumed experts are "correct[16]." As such, initial claims of expertise are not necessarily grounded in demonstrated judgmental ability, but are instead based on perceptive criteria such as credentials, experience, and nomination by peers.

[16] Especially in high-stakes, time-sensitive decision situations (e.g., military combat action, firefighting), "correct" does not necessarily mean optimal but rather something that at least adequately solves the problem at hand while minimising negative consequences as much as possible.

Taking this approach of expertise by assumption, it is possible that the expert assessors present in the studied samples were hired based on their perceived ability to make accurate judgments, but may not truly have been able to do so when put into actual practice. If a distinction must be made between assumed and true expertise, it then becomes more troubling that a prominent (but anonymous) management consulting firm offering services for high-stakes employment testing did not assign assessors who were truly experts in the first place.

A Criterion Problem

A possible reason why lower than average validities were observed in this dissertation may be that the supervisory ratings of job performance used as the criterion were deficient or otherwise inappropriate representations of the candidates' actual job performance. Such a "criterion problem" in identifying an adequate or the best representation of job performance has long plagued the field of industrial/organisational psychology (Austin & Villanova, 1992). Nowadays, a better representation of job performance is available in Campbell's (2012) eight-dimension model of job performance[17], which happen to parallel the assessment dimensions featured in this dissertation (albeit not a one-to-one correspondence). In practice however, it is very likely that rating measures of job performance are deficient with regard to these dimensions, contaminated with information not related to job performance, or fraught

[17] The dimensions are: 1) job-specific technical task proficiency, 2) non-job-specific technical task proficiency, 3) written and oral communication task proficiency, 4) demonstrating effort, 5) maintaining personal discipline, 6) facilitating peer and team performance, 7) supervision/leadership, and 8) management/administration.

with other sources of error (Campbell, 2012). To the extent that these supervisory ratings are poor measures of actual job performance, they would be more difficult to accurately predict if the goal is to predict actual job performance. It is worth noting that the ratings obtained here were for research rather than administrative purposes, which should improve their veracity.

The other side of the same criterion coin concerns what the expert assessors were actually trying to predict. Their official goal was to rate each candidate on person-job fit intended to eventually predict their performance once on the job. If they instead tried to predict other non-performance variables (e.g., likelihood of quitting the job), then their rating would be deficient with regard to job performance and contaminated with irrelevant information. Additionally, as observed in the relative weights analyses from Study 2, the assessors' ideas about what is most important for predicting performance differs from those of the supervisor raters. Therefore, it is possible that the expert assessors are more competent than the present findings have led us to believe, but also that their competence lies in predicting something they were not originally set out to predict. Future studies should examine whether experts are better at predicting certain types of criteria, such as performance-related versus non-performance-related or objective versus subjective.

Access to Additional Information

What all of the mechanical methods implemented in the analyses for this dissertation have in common is that they all combined the seven assessment dimensions

into an overall rating, albeit using different weighting schemes. Yet it is true that the expert assessors could have had more information about the candidates beyond scores on these seven assessment dimensions, such as their performance on individual assessment activities, test profiles, and biographical information obtained from sources such as résumés and personal interaction. This could be viewed as an advantage that a human judge has over a mechanical method. Despite the possibility that the experts had this information available to them, they still performed worse than any mechanical method. Prior research has shown that although people tend to become more confident about their judgments with more information available, they are not always more accurate (Tsai, Klayman, & Hastie, 2008).

If this use of additional information was actually the case, it may have ended up hindering the experts' judgments rather than being a blessing. As seen via the imperfect value for cue sensitivity in Study 1 and the relative weights analyses in Study 2 showing that the weighting structure of the assessor models were discrepant from the structure of the optimal ecology models, the expert assessors did not combine information in a way that reflected what was optimal in reality. It is possible that this discrepancy may have been, at least in part, due to this outside information influencing what the experts ended up concluding to be the most important assessment dimensions for evaluating their candidates. For example, they may have perceived a handful of candidates to have extraordinary leadership ability, which would end up making the leadership dimension more salient. As a result, they may have emphasised leadership in their evaluations when

in reality leadership contributed little to performance once on the job. Further research will be needed to test this hypothesis.

Data Limitations

There are a couple features (or shortcomings) of the data used in this dissertation that limited the scope of the analyses that could be conducted. These come down to issues of luck and opportunity in obtaining datasets with desired characteristics or qualities[18]. The first issue concerns the level of analysis. Ideally, the most powerful analysis would have been to model each individual assessor and test the predictive accuracy of each assessor. However, the relatively small sample sizes in the validation datasets meant that the "models of man" could only be confidently derived at the level of each dataset to reflect the average model across multiple assessors. Therefore, models were derived for Company A, Company B, Sample 1, and Company B, Sample 2, but not for the individual assessors in each sample. This sample size limitation is fairly characteristic of research in this domain as many Lens Model studies are conducted at an aggregated level (Karelaia & Hogarth, 2008).

The second data issue concerns the timeliness of the data as they are at least a couple decades old at the time of this writing. It is possible that the state of expert judgment in individual assessments has improved over time, but the fact that the stubborn reliance on expert judgment is still a recent subject for discussion (e.g., Highhouse, 2008;

[18] The data used in this dissertation has been a truly invaluable resource and has made a plethora of analyses possible. As a researcher and generally optimistic human being however, I would like to believe that we can always do better.

Kuncel & Highhouse, 2011) and that recent studies still find that people prefer subjective over mechanical methods of judgment (Diab et al., 2011; Eastwood et al., 2012) suggests that this problem still persists in the present day. Furthermore, time does not necessarily heal. In this dissertation for example, Company B, Sample 2 was a validation study conducted after the Company B, Sample 1 validation study, but validities were lower across the board for Sample 2 compared to Sample 1. More information not currently available in these datasets would be needed to determine whether there are factors other than time at play, but it does seem that past lessons have not been completely effective at informing future practices. These issues of subjective versus mechanical judgment have been discussed as early as Meehl (1954), so it is perhaps a bit disheartening to find that the basic points still need to be argued today in the face of overwhelming evidence demonstrating the general inefficacy of human or expert judgment, as well as evidence in support of the superiority of mechanical methods of judgment.

Practical Recommendations

Even though there has been a large body of evidence for the superiority of mechanical methods of judgment over human judgment across a range of decision contexts, I have no intention of suggesting that human or expert judgment should be replaced by purely mechanical methods. There is enough resistance against it and preference towards expert judgment (Diab et al., 2011; Eastwood et al., 2012, Highhouse, 2008; Kleinmuntz, 1990) that it would not be a realistic option. I believe that there is

enough face[19] validity in expert judgment that makes it worth retaining. That said, if expert judgment is to be retained, it clearly needs to be improved. Fortunately, there are options for supporting expert judgment with mechanical methods.

"Model of man." Evident from the imperfect values of cognitive control in Study 1, along with the finding that *mindless consistency* is enough to outperform the expert assessors in Study 3, one of the main concerns that detriments the validity of expert judgment is the inconsistency with which the expert assessors apply their weighting schemes when evaluating multiple candidates. By modeling the assessors (i.e., subject side of the Lens Model), we can obtain the average judgment policy for an assessor or across multiple assessors, and then apply this assessor model consistently across all cases as was done in Studies 1 and 2. Studies 2 and 3 also showed that while it is best to have optimal weights, the actual weights used in a model really does not matter as much as the fact that the weights are being applied consistently. As mentioned previously, linear models are robust to changes in the weights as long as they do not flip signs (Dawes, 1979). By modeling the assessors and then applying their model to evaluate candidates, expert judgment is retained in terms of specifying the model, and it ensures the consistency with which the expert's policy is applied. This procedure was tested as early as Goldberg (1970), and works just as well today.

Just "man." Since the optimality of the weighting scheme matters less than the consistency with which it is applied, the simpler option would be to just ask an expert assessor to define weighting scheme and then apply that expert-defined scheme

[19] Quite literally, an expert judge has a face whereas a mechanical model does not.

mechanically. To wash out individual error in defining a set of weights, it would be even better to take the average across a group of expert-defined weighting schemes. Functionally, this can be applied exactly the same as the "model of man" method, except that it skips the modeling step. This is particularly beneficial when no judgment information has yet been collected, preventing the "model of man" from being derived.

Anchoring. If expert input is desired in the actual judgmental data combination process and not simply the initial definition of weights, judgmental consistency could be improved via anchoring. When people are provided anchor values, they typically do not adjust their judgmental process too far from the anchor value (Hastie & Dawes, 2001; Tversky & Kahneman, 1974). In the present case, the "model of man" dimension weights can be provided as an anchor weighting scheme, or the expert assessor could also be asked to define their own weighting scheme. By having an explicit set of anchor dimension weights defined and available to the assessor, it may reduce the assessors' volatility in weighting the assessment dimensions across multiple candidates. It may also be effective to set limits on adjustments away from the anchors or require justification for adjusting away from the anchors.

Clinical and mechanical synthesis. Clinical and mechanical synthesis were described by Sawyer (1966). In clinical synthesis, the assessor is given the mechanical composite rating, and is then free to make the final prediction by combining the mechanical composite with the original data. In mechanical synthesis, the assessor combines the dimension information using their own judgmental policy, after which the expert's judgment is mechanically combined with the original data to obtain a final

overall judgment. Sawyer (1966) showed that when the assessor has final say in clinical synthesis, validity was better than if the evaluation was purely left up to the assessor, but was worse than the purely mechanical composite. With mechanical synthesis, including the judge's clinical prediction in the final prediction still preserves predictive accuracy relative to a purely mechanical composite that excludes the clinical prediction. Both clinical and mechanical synthesis would be expected to result in better prediction than the assessor alone, but deciding between the two would be a trade-off between the amount of control given to the assessor and the level of accuracy that could be achieved.

Limit expert judgment. If the goal is to select the best candidate or a few top candidates from a larger candidate pool (i.e., if the selection ratio is low), candidates can be pre-screened using mechanical methods of judgment, after which the assessor can select among the top candidates (Kuncel, 2008). This limits the use of expert judgment to a set of candidates with similar predicted performance scores, and they are likely to be fairly invariant in terms of their actual job performance. Even if the expert does not select the optimal choice, the loss in performance (or other desired outcomes) by selecting a candidate a few ranks lower is likely to be negligible.

Aggregate expert judgment. An additional procedure that could be carried out with any of the aforementioned procedures is to aggregate the judgments of multiple assessors. The goal here is to basically wash out any individual errors in judgment and promote any judgmental policies that are in agreement, which should result in more consistent judgments across cases.

There are two ways of approaching the aggregation of expert judgment. One way is to have the expert assessors work in a group to come up with an overall group judgment. Such group decision making has been shown to outperform the individuals (e.g., Sniezek & Henry, 1989), but the process gains from group decision making may be offset by process losses (e.g., groupthink, social loafing) that introduce new forms of error (Kerr & Tindale, 2004). The other way of aggregating expert judgment is to simply have the expert assessors each make their judgments individually, and then take the average of their individual judgments. Here, process losses from active group decision making are avoided at the expense of missing out on possible gains above minimising individual error. With this method, it is also possible to differentially weight the contributions of each individual assessor if doing so could be reasonably expected to improve predictive accuracy. For example, the judgments of more experienced assessors or assessors who have demonstrated high accuracy in the past could be weighed more heavily than those less experienced or those with a poorer track record.

Conclusion

Even if we have the good fortune of possessing a set of predictors that perfectly relates to the criterion, their predictive power is ultimately dependent on how that predictor information is used. Validity will be negatively impacted if the data combination method introduces error instead of maximising the predictive power of those predictors. When using individual assessments to evaluate candidates for jobs, the longstanding reliance on expert judgment has been a significant hindrance to ensuring the validity of these assessments.

The bad news is that expert assessors have frequently been found to be both suboptimal and inconsistent in their use of predictor weighting schemes to combine assessment dimension information, resulting in lower levels of validity than what could have theoretically been achieved. Expertise is in no way a guarantee of validity, and attempts to exercise expert insight often ends up introducing more error than validity. The good news is that there are methods that can potentially reduce human error and improve the consistency of expert judgment while still retaining the use of expert judgment so as to avoid negative reactions toward the pure use of mechanical methods. Further research into these methods and continual development of new methods of improving human judgment will not only improve our ability to maximise the predictive and face validity of our assessments, but can also be more broadly applied to improving judgment and decision making processes, no matter the context.

References

Arthur, W., Day, E. A., McNelly, T. L., & Edens, P. S. (2003). A meta-analysis of the criterion-related validity of assessment center dimensions. *Personnel Psychology, 56,* 125-153.

Astin, A.W. (1961). The functional autonomy of psychotherapy. *American Psychologist, 16,* 75-78.

Austin, J. T., & Villanova, P. (1992). The criterion problem: 1917-1992. *Journal of Applied Psychology, 77,* 836-784.

Bobko, P., Roth, P. L., & Potosky, D. (1999). Derivation and implications of a meta-analytic matrix incorporating cognitive ability, alternative predictors, and job performance. *Personnel Psychology, 52,* 561-589.

Brunswik, E. (1952). *The conceptual framework of psychology.* Chicago: University of Chicago Press.

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review, 62,* 193–217.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments* (2nd ed.). Berkeley, CA: University of California Press.

Camerer, C. F., & Johnson, E. J. (1991). The process-performance paradox in expert judgment: How can experts know so much and predict so badly? In K. A. Ericsson & J. Smith (Eds.), *Toward a general theory of expertise: Prospects and limits* (pp. 195–217). Cambridge, UK: Cambridge University Press.

Campbell, J. P. (2012). Behavior, performance, and effectiveness in the twenty-first century. In S. W. J. Kozlowski (Ed.), *The Oxford handbook of organizational psychology, Vol. 1* (pp. 159-196). New York, NY: Oxford University Press.

Cooksey, R. W. (1996). *Judgment analysis: Theory, methods, and applications.* San Diego, CA: Academic Press.

Dawes, R. M. (1971). A case study of graduate admissions: Application of three principles of human decision making. *American Psychologist, 26,* 180–188.

Dawes, R. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34,* 571–582.

Dawes, R., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin, 81,* 95–106.

De Corte, W., Lievens, F., & Sackett, P. R. (2007). Combining predictors to achieve optimal trade-offs between selection quality and adverse impact. *Journal of Applied Psychology, 92,* 1380.

Diab, D. L., Pui, S., Yankelevich, M., & Highhouse, S. (2011). Lay perceptions of selection decision aids in US and non-US samples. *International Journal of Selection and Assessment, 19,* 209-216.

Eastwood, J., Snook, B., & Luther, K. (2012). What people want from their professionals: Attitudes toward decision-making strategies. *Journal of Behavioral Decision Making, 25,* 458-568.

Einhorn, H. J. (1974). Cue definition and residual judgment. *Organizational Behavior and Human Performance, 12,* 30-49.

Garb, H. N. (1989). Clinical judgment, clinical training, and professional experience. *Psychological Bulletin, 105,* 387-396.

Goldberg, L. R. (1965). Diagnosticians vs. diagnostic signs: The diagnosis of psychosis vs. neurosis from the MMPI. *Psychological Monographs,* 79 (9, Whole No. 602).

Goldberg, L. R. (1970). Man versus model of man: A rationale, plus some evidence, for a method of improving on clinical inference. *Psychological Bulletin, 73,* 422–432.

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. *Psychology, Public Policy, & Law, 2,* 293–323.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction. *Psychological Assessment, 12,* 19–30.

Hammond, K. R. (1955). Probabilistic functioning and the clinical method. *Psychological Review, 62,* 255–262.

Hammond, K. R., Hursch, C. J., & Todd, F. J. (1964). Analyzing the components of clinical inference. *Psychological Review, 71,* 438–456.

Hastie, R., & Dawes, R. M. (2001). *Rational choice in an uncertain world.* Thousand Oaks, CA: Sage.

Hazucha, J.; Ramesh, A.; Goff, M.; Crandell, S.; Gerstner, C.; Sloan, E.; Bank, J.; & van Katwyk, P. (2011). Individual psychological assessment: The poster child of blended science and practice. *Industrial and Organizational Psychology, 4,* 298-302.

Highhouse, S. (1997). Understanding and improving job-finalist choice: The relevance of behavioral decision research. *Human Resource Mangagement Review, 7,* 449-470.

Highhouse, S. (2002). Assessing the candidate as a whole: A historical and critical analysis of individual psychological assessment for personnel decision making. *Personnel Psychology, 55,* 363-396.

Highhouse, S. (2008). Stubborn reliance on intuition and subjectivity in employee selection. *Industrial and Organizational Psychology, 1,* 333–342.

Hoffman, P. J. (1960). The paramorphic representation of clinical judgment. *Psychological Bulletin, 57,* 116-131.

Humphrey, S. E., Nahrgang, J. D., & Morgeson, F. P. (2007). Integrating motivational, social, and contextual work design features: A meta-analytic summary and theoretical extension of the work design literature. *Journal of Applied Psychology, 92,* 1332-1356.

Hursch, C. J., Hammond, K. R., & Hursch, J. L. (1964). Some methodological considerations in multiple-probability studies. *Psychological Review, 71,* 42-60.

Jeanneret, R., & Silzer, R. (1998). An overview of individual psychological assessment. In R. Jeanneret & R. Silzer (Eds.), *Individual psychological assessment* (pp. 3-26). San Francisco, CA: Jossey-Bass.

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist, 64,* 515-526.

Karelaia, N., & Hogarth, R. M. (2008). Determinants of linear judgment: A meta-analysis of lens model studies. *Psychological Bulletin, 134,* 404-426.

Kerr, N. L., & Tindale, R. S. (2004). Group performance and decision making. *Annual Review of Psychology, 55,* 623-655.

Klein, G. (2008). Naturalistic decision making. *Human Factors, 50,* 456-460.

Klein, G. A., Calderwood, R., & Clinton-Cirocco, A. (1986). Rapid decision making on the fireground. *Proceedings of the Human Factors and Ergonomics Society 30th Annual Meeting, 1,* 576–580.

Klein, G. A., Orasanu, J., Calderwood, R., & Zsambok, C. E. (1993). *Decision making in action: Models and methods.* Norwood, NJ: Ablex.

Kleinmuntz, B. (1990). Why we still use our heads instead of formulas: Toward an integrative approach. *Psychological Bulletin, 107,* 296–310.

Klimoski, R. J., & Zukin, L. B. (2003). Psychological assessment in industrial/organizational settings. In J. R.Graham & J. A.Naglierie (Eds.), *Assessment psychology* (pp. 317–343). Hoboken, NJ: Wiley.

Kristof-Brown, A. L., Zimmerman, R. D., & Johnson, E. C. (2005). Consequences of individuals' fit at work: A meta-analysis of person-job, person-organization, person-group, and person-supervisor fit. *Personnel Psychology, 58,* 281-342.

Kuncel, N. R. (2008). Some new (and old) suggestions for improving personnel selection. *Industrial and Organizational Psychology, 1,* 343-346.

Kuncel, N. R., & Highhouse, S. (2011). Complex predictions and assessor mystique. *Industrial and Organizational Psychology, 4,* 302-306.

Kuncel, N. R., Klieger, D. M., Connelly B. S., & Ones, D.S. (2013). Mechanical versus clinical data combination in selection and admissions decisions: A meta-analysis. *Journal of Applied Psychology, 98,* 1060-1072.

Kwaske, I. (2004). Individual assessments for personnel selection: An update on a rarely researched but avidly practiced practice. *Consulting Psychology Journal: Practice and Research. 56,* 186-195.

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence.* Minneapolis, MN: University of Minnesota.

Morris, S., Daisley, R., Wheeler, M., & Boyer, P. (2015). A meta-analysis of the relationships between individual assessment and job performance. *Journal of Applied Psychology, 100,* 5-20.

Nisbett, R. E., Zukier, H., & Lemley, R. E. (1981). The dilution effect: Nondiagnostic information weakens the implications of diagnostic information. *Cognitive Psychology, 13,* 248-277.

Phillips, J. K., Klein, G., & Sieck, W. R. (2004). Expertise in judgment and decision making: A case for training intuitive decision skills. In D. J. Koehler, & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 297-315). Oxford, UK: Blackwell.

Ryan, A. M., & Sackett, P. R. (1987). A survey of individual assessment practices by I/O psychologists. *Personnel Psychology, 40,* 455-488.

Sackett, P. R., Gruys, M. L., & Ellingson, J. E. (1998). Ability-personality interactions when predicting job performance. *Journal of Applied Psychology, 83,* 545-556.

Sawyer, J. (1966). Measurement and prediction, clinical and statistical. *Psychological Bulletin, 66,* 178–200.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124,* 262-274.

Silzer, R., & Jeanneret, R. (2011). Individual psychological assessment: A practice and science in search of common ground. *Industrial and Organiational Psychology, 4,* 270-296.

Simon, H. A. (1992). What is an explanation of behavior? *Psychological Science, 3,* 150-161.

Sniezek, J. A., & Henry, R. A. (1989). Accuracy and confidence in group judgment. *Organizational Behavior and Human Decision Processes, 43,* 1-28.

Society for Industrial and Organizational Psychology, Inc. (2016). *Guidelines for education and training in industrial-organizational psychology.* Bowling Green, OH: Author.

Spengler, P. M., White, M. J., Ægisdóttir, S., Maugherman, A. S., Anderson, L. A., Cook, R. S., ... & Rush, J. D. (2009). The meta-analysis of clinical judgment project: Effects of experience on judgment accuracy. *The Counseling Psychologist, 37,* 350-399.

Spychalski, A. C., Quiñones, M. A., Gaugler, B. B., & Pohley, K. (1997). A survey of assessment center practices in organizations in the United States. *Personnel Psychology, 50,* 71-90.

Thorndike, E. L. (1918). Fundamental theorems in judging men. *Journal of Applied Psychology, 2,* 67-76.

Thornton, G.C., Hollenbeck, G.P., Johnson, S.K. (2010). Selecting leaders: Executives and high potentials. In J.L. Farr, N.T. Tippins (Eds.) *Handbook of employee selection* (pp. 823-840). New York: Taylor & Francis.

Tsai, C. I., Klayman, J., & Hastie, R. (2008). Effects of amount of information on judgment accuracy and confidence. *Organizational Behavior and Human Decision Processes, 107,* 97-105.

Tucker, L. R. (1964). A suggested alternative formulation in the developments by Hursch, Hammond and Hursch and by Hammond, Hursch and Todd. *Psychological Review, 71,* 528-530.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185,* 1124–1131.

Tversky, A., & Kahneman, D. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47,* 263-291.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty, 5,* 297-323.

Waller, N. G. (2008). Fungible weights in multiple regression. *Psychometrika, 73,* 691-703.

Table 1
Summary of Studies

Study 1: A Lens Model Decomposition of Individual Assessments	
Purpose	To break down expert judgment in individual assessments into its component processes, and to evaluate their contributions (or hindrances) toward the use of individual assessments to predict job performance in comparison to mechanical data combination methods.
Sample	Archival individual assessment validation datasets: - Expert-judged overall assessment ratings for each candidate based on ratings on seven assessment dimensions (adjustment, administration, communication, interpersonal, judgment, leadership, and motivation), which were based on each candidate's performance on a mix of in-basket, interviews, leaderless group discussions, personality tests, and cognitive ability tests - Supervisory ratings of job performance 1) Company A (financial services provider) - 231 candidates - 26 assessors - Conducted between 1994 and 1997 2) Company B (food retailer), Sample 1 - 195 candidates - 23 assessors - Conducted between 1980 and 1988 3) Company B, Sample 2 - 421 candidates - 30 assessors - Conducted between 1989 and 1999
Analyses	- Lens Model parameters calculated as described in Table 2 - Regression modeling via ordinary least squares multiple regression - Analyses conducted at the level of each dataset, then additionally aggregated via sample-size weighting
Hypotheses	- Mechanical data combinations will outperform clinical combinations (expert judgment) in predicting job performance - Bootstrapped "model of man" will outperform expert judgment - Low unmodeled knowledge: observed criterion values and subject judgments are sufficiently modelled by multiple linear regression - Moderate to high cue sensitivity: judges use cue weights that are not perfectly inconsistent with ecological weights in either or both magnitude and rank order - Moderate to high cognitive control: judges are not perfectly consistent in their application of a single cue weighting policy across judgments

Table 1
Summary of Studies - continued

Study 2: Local versus Non-Local Models of Expert Judgment	
Purpose	To test the validity of expert insight by evaluating whether models of expert assessors derived from a specific organisation are better at predicting job performance at the same organisation compared to models of expert assessors derived from a different organisation.
Sample	Archival individual assessment validation datasets: - Same as Study 1 Archival general individual assessment dataset: - Does not contain job performance validation information - 16,143 candidates - 176 assessors - 683 organisations - Conducted between 1971 and 2000

Analyses	- "Model of man" derived for each dataset (expert-judged overall assessment ratings regressed on dimension ratings)
	- Each model applied to each validation dataset (all possible combinations) to combine assessment dimensions ratings into "model of man" overall assessment ratings
	- Other mechanical composites of dimension ratings obtained via:
	  - Unit weighting via simple sums (sum of all dimensions)
	  - Optimal regression weighting
	- Predictive validity of these methods of obtaining the overall assessment ratings quantified by correlating them with supervisory ratings of job performance

Hypotheses	- Mechanical methods will outperform expert judgment
	- Overall ratings made via optimal regression weighting will perform the best
	- If expert insight is valid:
	  - The model local to one specific organisational sample should outperform the model from a different organisational sample or from a model not specific to any one organisation (i.e., the general model)
	- If expert insight is not valid:
	  - The model local to one specific organisational sample would not outperform the model from a different organisational sample or from a model not specific to any one organisation (i.e., the general model)
	  - Non-local model may outperform the local model by chance

Table 1
Summary of Studies - continued

Study 3: Comparing Random Weighting Schemes with Expert Judgments	
Purpose	To determine the extent to which consistency in applying a single predictor weighting scheme contributes to predictive validity by testing random weighting methods to isolate consistent weighting from accurate weighting.
Sample	Archival individual assessment validation datasets: - Same as Study 1
Analyses	Monte Carlo simulation: 1) Random weights applied consistently - Seven weights randomly sampled from a uniform distribution ranging from 0 to 0.5, inclusive - This same set of weights used to linearly combine each candidate's dimension ratings into an overall assessment rating 2) Random weights applied randomly - Seven weights randomly sampled from a uniform distribution ranging from 0 to 0.5, inclusive - A new set of random weights generated to linearly combine each candidate's dimension ratings into an overall rating - Predictive validity of each method quantified by correlating the overall assessment ratings with supervisory ratings of job performance - 10,000 iterations for each method, for a total of 10,000 validity coefficients each - Compared to non-random methods of data combination: - Expert judgment - Simple sums - Optimal regression weighting
Hypotheses	- Mechanical methods will outperform expert judgment - Optimal regression weighting will perform the best on average - Non-random methods will outperform random methods on average - Consistent random weights will outperform completely random weights on average - Consistent random weights will outperform expert judgment in an overwhelming majority of iterations - If judges are not consistent with using a single weighting policy and if this inconsistency does not reflect valid expert insight, completely random weights could potentially match or outperform expert judgment

Table 2
Lens Model Parameters

Parameter	Description	Formula
Collected Parameters		
Y_e	Observed criterion values	
Y_s	Subject's (judge's) judgments	
$X_1 \ldots X_n$	Values of n independent variable cues (i.e., predictors, indicators)	
Modelled Parameters		
$b_{1e} \ldots b_{ne}$	Ecological model cue weights obtained by regressing the observed criterion values on the cues. These are the optimal weights for predicting the criterion from the cues; the magnitude of each cue weight describes its relative strength in predicting the criterion.	$\hat{Y}_e = b_{1e}X_1 + \cdots + b_{ne}X_n$
$b_{1s} \ldots b_{ns}$	Subject model cue weights obtained by regressing the subject judgments on the cues. These are "model of man" weights that capture the subject's data combination policy for these cues; the magnitude of each cue weight describes the relative importance the subject places on each cue in making the judgment.	$\hat{Y}_s = b_{1s}X_1 + \cdots + b_{ns}X_n$
$\hat{Y}_e$	Criterion values predicted from the ecological model. Cue values mechanically combined using the ecological model weights.	$\hat{Y}_e = b_{1e}X_1 + \cdots + b_{ne}X_n$
$\hat{Y}_s$	Subject's judgments predicted from the subject model. Cue values mechanically combined using the subject model weights. Also known as a bootstrapped "model of man" prediction.	$\hat{Y}_s = b_{1s}X_1 + \cdots + b_{ns}X_n$
$Y_{e_{res}}$	Residual term for the ecological model; difference between the predicted and observed criterion values.	$Y_{e_{res}} = Y_e - \hat{Y}_e$
$Y_{s_{res}}$	Residual term for the subject model; difference between the predicted and observed subject judgments.	$Y_{s_{res}} = Y_s - \hat{Y}_s$

Table 2
Lens Model Parameters - continued

Parameter	Description	Formula
Computed Parameters		
r_a	Accuracy (i.e., clinical validity, achievement index). Correlation between the observed criterion values and the subject judgments; the criterion-related validity of the subject judgments in predicting the criterion values.	$r_a = r_{Y_e Y_s}$ $= GR_eR_s + C\sqrt{(1-R_e^2)(1-R_s^2)}$
R_e	Environmental predictability. Correlation between the observed criterion values and the ecological model-predicted criterion values; the criterion-related validity of the optimally weighted mechanical combination of cues in predicting the criterion values.	$R_e = r_{Y_e \hat{Y}_e}$
R_s	Cognitive control. Correlation between the subject judgments and the "model of man"-predicted subject judgments; the criterion-related validity of the "model of man" in predicting the subject judgments.	$R_s = r_{Y_s \hat{Y}_s}$
G	Cue sensitivity (i.e., matching index, mechanical knowledge). Correlation between the predicted criterion values and predicted subject judgments from their respective models; captures the degree to which the "model of man" weights are consistent with the optimal weighting scheme.	$G = r_{\hat{Y}_e \hat{Y}_s}$
C	Unmodeled knowledge (i.e., residual correlation). Correlation between the residuals from each model; captures any relationship between the observed criterion and subject judgment that may be attributed to random error as well as systematic error variance due to model misspecification: unmodeled cues, unmodeled functional relationships (e.g., nonlinear relationships if the model is linear), and unmodeled interactions.	$C = r_{Y_{e_{res}} Y_{s_{res}}}$

Table 2
Lens Model Parameters - continued

Parameter	Description	Formula
Parameter Composites		
GR_e	Correlation between the observed criterion values and the bootstrapped "model of man" predicted judgments; criterion-related validity of the "model of man" mechanical composite in predicting the criterion.	$GR_e = r_{Y_e\hat{Y}_s}$
GR_s	Correlation between the subject judgments and the ecological model-predicted criterion values; captures the degree to which the subject applies a cue weighting scheme consistent with the ecological model, and applies it consistently across judgments.	$GR_s = r_{Y_s\hat{Y}_e}$
$GR_e - r_a$	Difference between the predictive validity of the subject model and subject judgments in predicting the criterion; quantifies how much better (or worse, if negative) the judge would have done in prediction if he or she had applied their average data combination policy consistently across judgments.	
GR_eR_s	Mechanical component of judgmental accuracy. If there is no unmodeled knowledge (i.e., $C = 0$), then $r_a = GR_eR_s$, meaning that accuracy is dependent only on how well the ecological and subject models match up, and the degree to which the criterion and subject judgments are predictable by the cues.	
$C\sqrt{(1 - R_e^2)(1 - R_s^2)}$	Unmodeled component of judgmental accuracy; quantifies the degree to which the unmodeled knowledge contributes to judgmental accuracy.	

Table 3
Ecology Models

Dataset	Dimension	b	SE	R^2
Company A	Motivation	.10	.14	.06
	Judgment	.05	.10	
	Administrative	.19	.10	
	Communication	-.32	.19	
	Interpersonal	.16	.20	
	Leadership	.14	.18	
	Adjustment	.07	.19	
Company B, Sample 1	Motivation	.30	.13	.17
	Judgment	.23	.08	
	Administrative	.21	.11	
	Communication	-.01	.13	
	Interpersonal	-.02	.11	
	Leadership	.04	.11	
	Adjustment	.06	.12	
Company B, Sample 2	Motivation	.10	.09	.09
	Judgment	-.02	.06	
	Administrative	.43	.11	
	Communication	-.04	.12	
	Interpersonal	.13	.09	
	Leadership	.02	.09	
	Adjustment	.00	.11	

Table 4
Subject Models

Dataset	Dimension	b	SE	R^2
Company A	Motivation	.01	.05	.69
	Judgment	.21	.03	
	Administrative	.04	.03	
	Communication	.10	.06	
	Interpersonal	.20	.06	
	Leadership	.34	.06	
	Adjustment	.13	.06	
Company B, Sample 1	Motivation	.27	.11	.54
	Judgment	.26	.09	
	Administrative	.15	.12	
	Communication	.08	.11	
	Interpersonal	.21	.10	
	Leadership	.10	.09	
	Adjustment	.29	.11	
Company B, Sample 2	Motivation	.04	.07	.57
	Judgment	.28	.03	
	Administrative	-.05	.07	
	Communication	.05	.08	
	Interpersonal	.21	.05	
	Leadership	.27	.05	
	Adjustment	.14	.06	
General Assessment Dataset	Motivation	.09	.01	.69
	Judgment	.22	.01	
	Administrative	.13	.01	
	Communication	.11	.01	
	Interpersonal	.20	.01	
	Leadership	.24	.01	
	Adjustment	.21	.01	

Table 5
Obtained Lens Model Parameters

Source	r_a	R_e	R_s	G	C	GR_e	GR_s	$GR_e - r_a$	GR_eR_s	$C\sqrt{(1-R_e^2)(1-R_s^2)}$
Company A	.17	.25	.83	.68	.05	.17	.56	.00	.14	.03
Company B, Sample 1	.20	.41	.73	.89	-.11	.36	.65	.16	.26	-.07
Company B, Sample 2	.13	.30	.75	.65	-.02	.20	.49	.06	.15	-.01
Sample Size Weighted Average	.16	.31	.77	.71	-.02	.23	.55	.07	.17	-.01
Karelaia & Hogarth (2008)	.56	.81	.80	.80	.04	.65	.66	.10		
Kuncel et al. (2013)	.28	.44								

Note. Refer to Table 2 for comprehensive descriptions of each Lens Model parameter.

Table 6

Predictive Validities of Different Data Combination Methods

		Assessor Models				"Sterile" Mechanical Methods		
Source	Expert Judgment	Company A	Company B, Sample 1	Company B, Sample 2	General Assessment Dataset	Simple Unit Weights	Optimal Weights	Optimal Weights (Adjusted)
Company A	.17	.17	.19	.15	.18	.19	.25	.18
Company B, Sample 1	.20	.31	.36	.31	.34	.35	.41	.37
Company B, Sample 2	.13	.22	.23	.20	.23	.24	.30	.28

Note. The values for "Optimal Weights (Adjusted)" are the validities of optimal weights adjusted to the population level via the Wherry formula-1 (Yin & Fan, 2001).

Table 7
Example Consistent and Completely Random Weighting Schemes

	Consistent Random Weights				Completely Random Weights			
Candidate	W1	W2	…	W7	W1	W2	…	W7
1	.02	.36		.19	.35	.45		.12
2	.02	.36		.19	.15	.11		.28
3	.02	.36	…	.19	.43	.06	…	.33
4	.02	.36		.19	.04	.41		.31
5	.02	.36		.19	.22	.17		.09

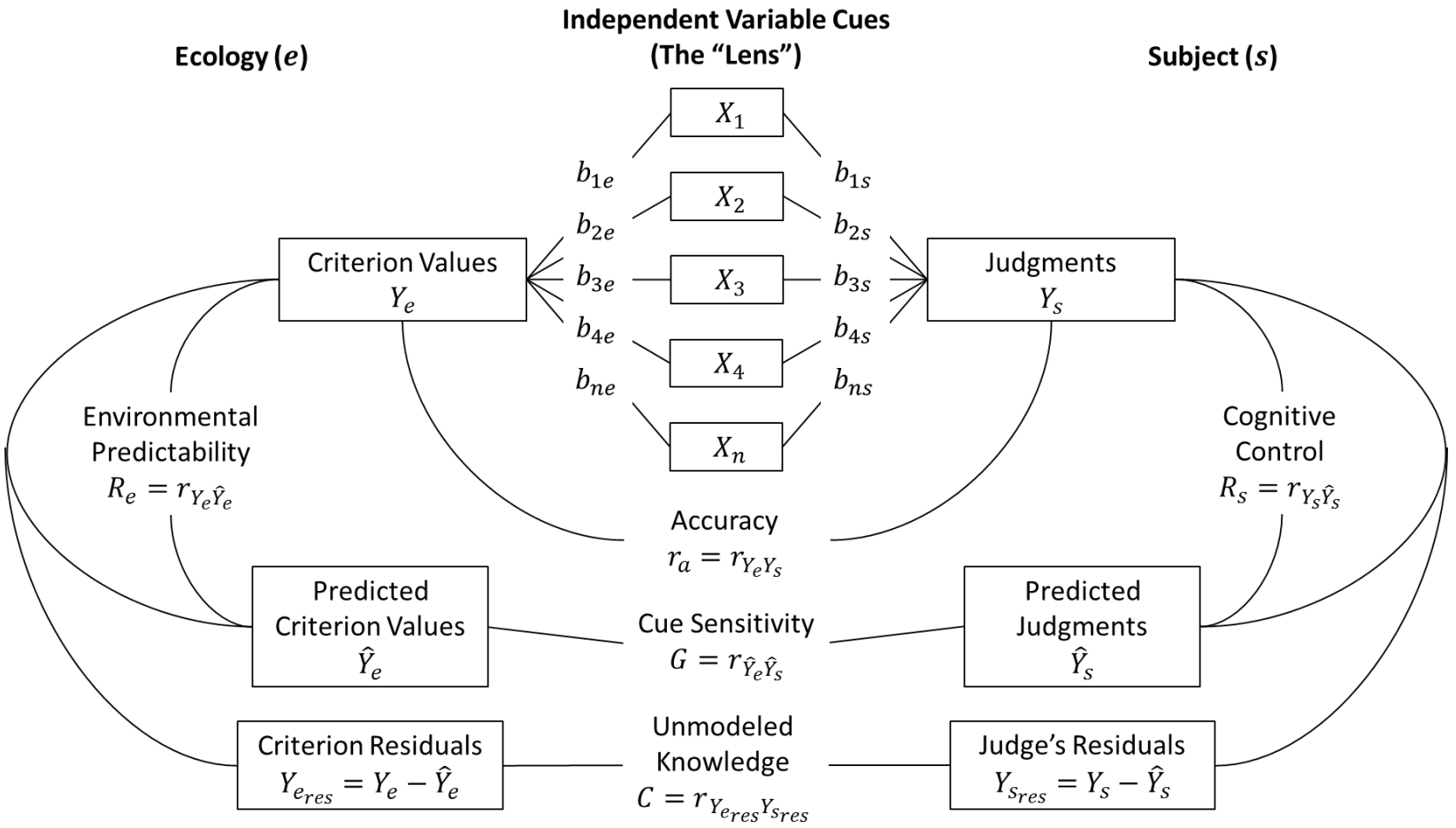

Figure 1. The Lens Model.

Note. Adapted with permission from Kuncel (personal communication).

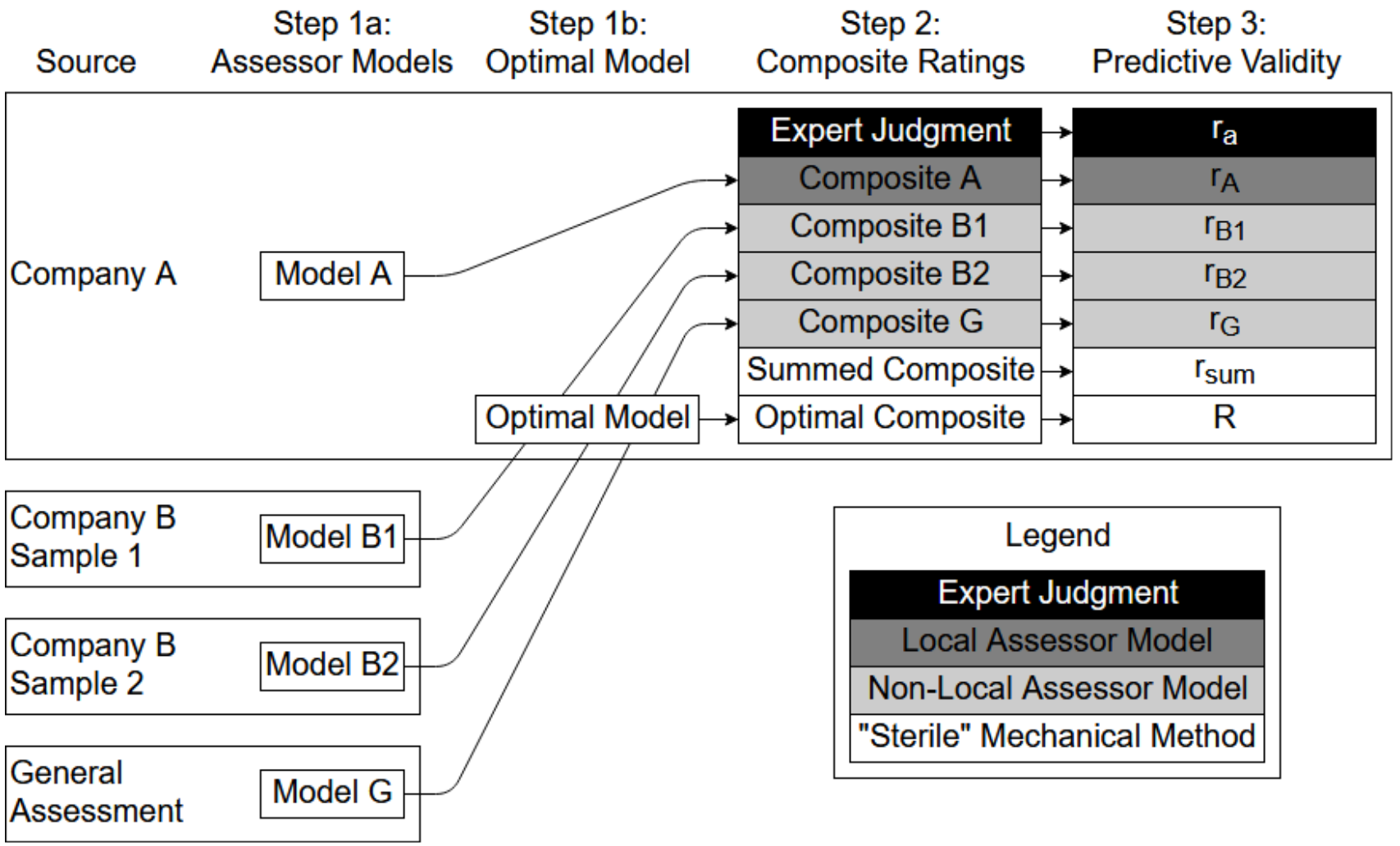

Figure 2. Example analytical plan for Study 2 for Company A. The same analysis is conducted for Company B, Sample 1, and Company B, Sample 2, except the local and non-local assessor models will be specific to each source dataset.

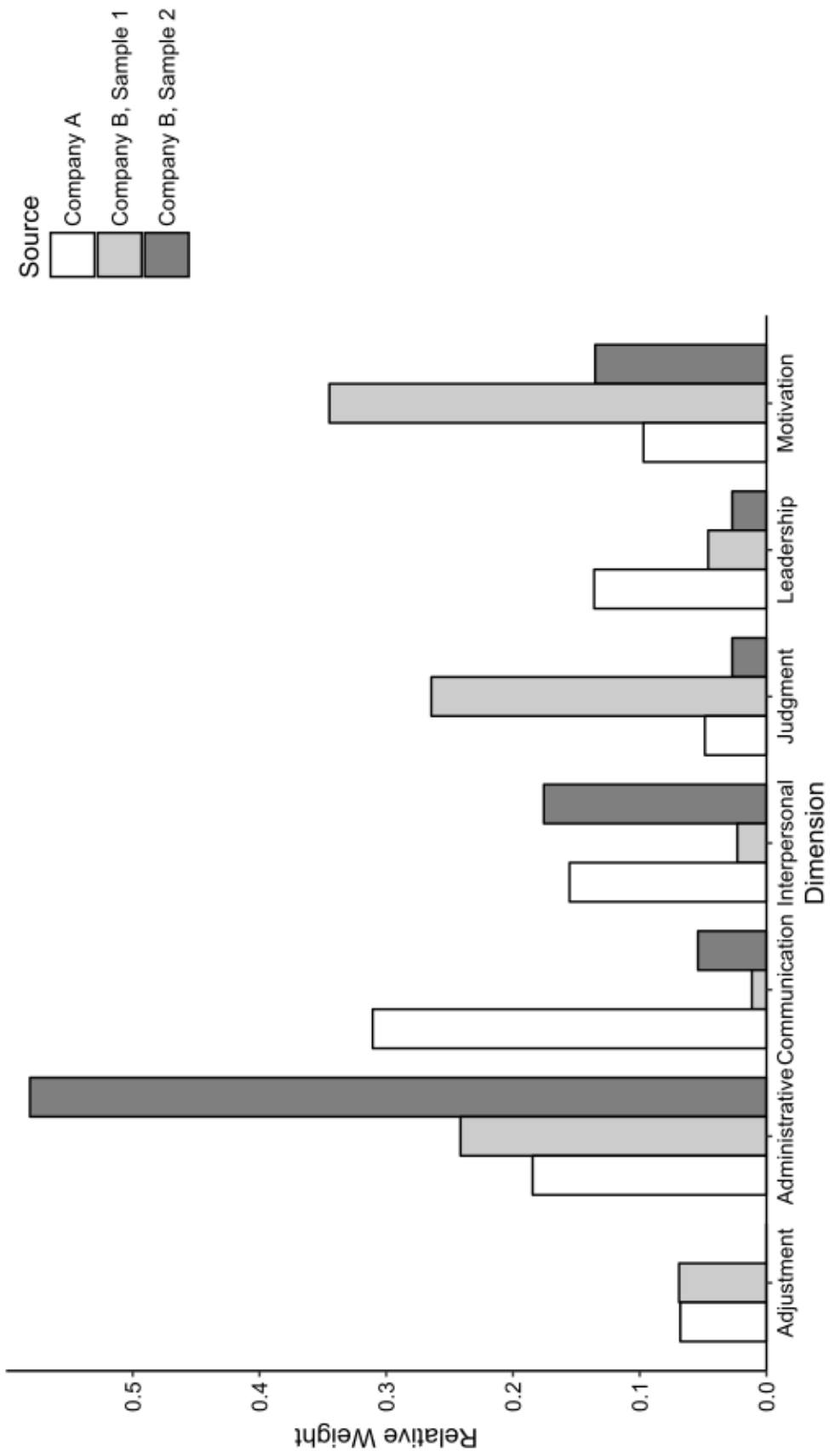

Figure 3. Relative weights for the ecology models.

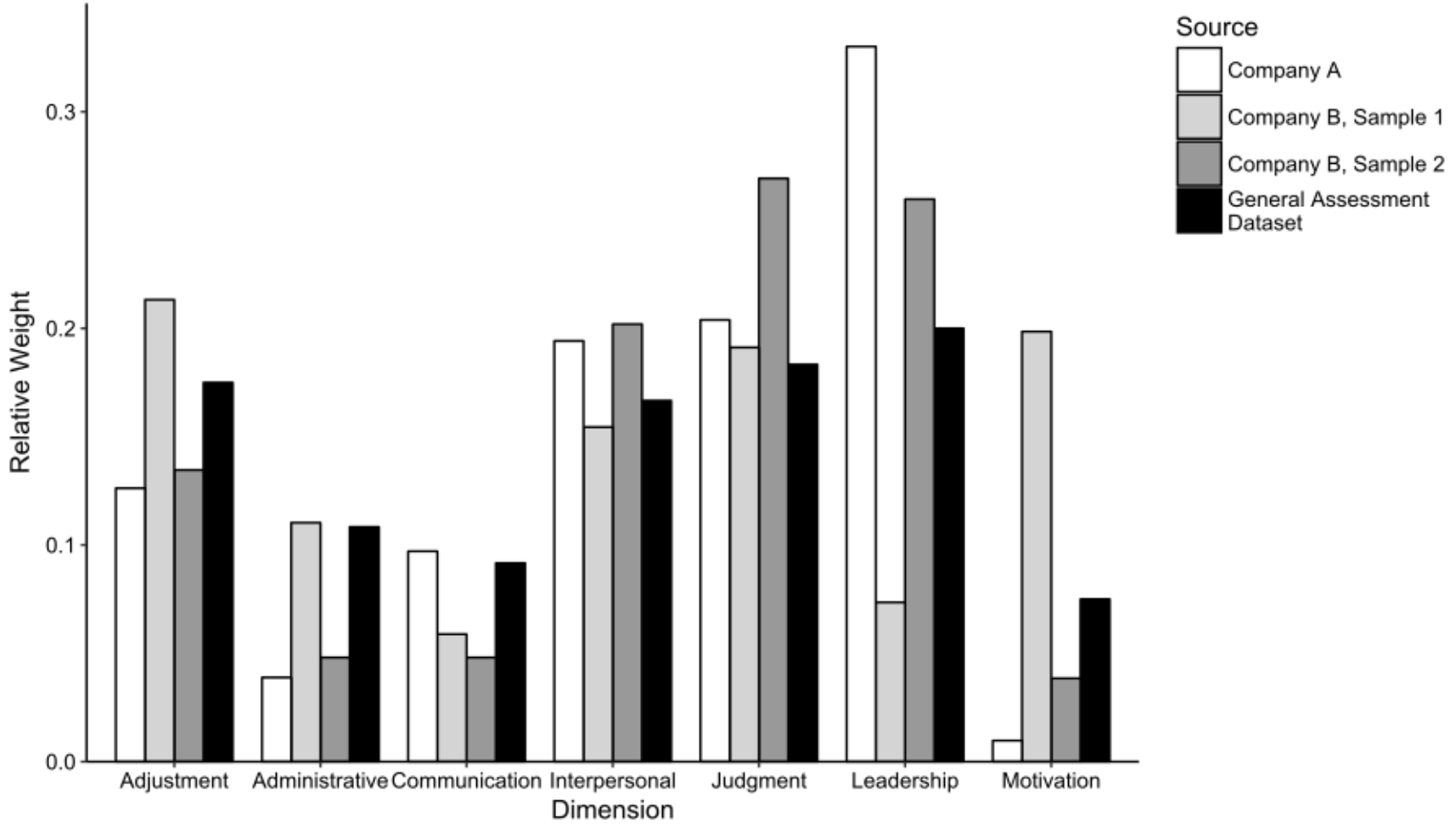

Figure 4. Relative weights for the subject models.

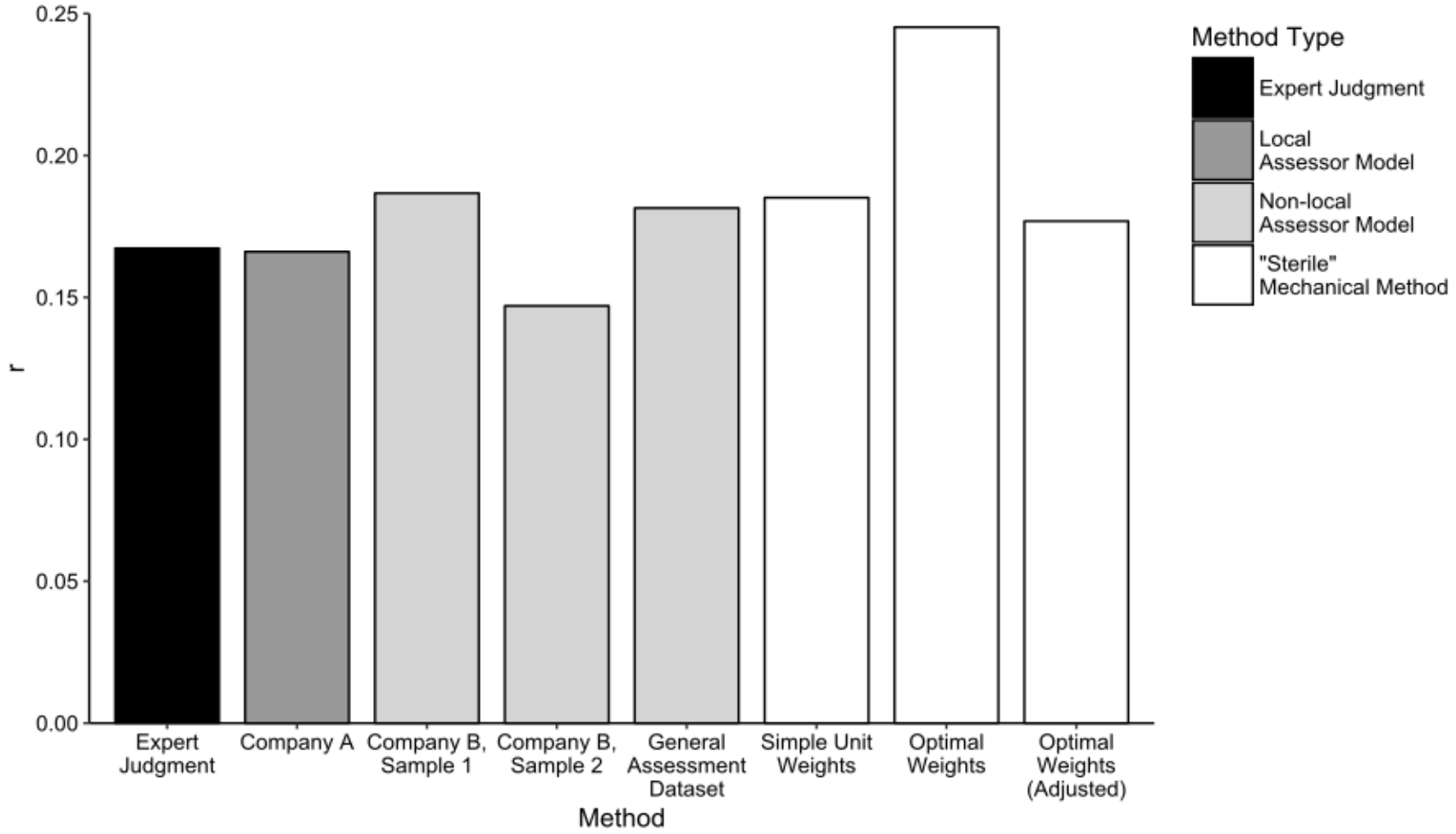

Figure 5. Validities of overall assessment ratings derived from different data combination methods for predicting supervisory ratings of job performance at Company A.

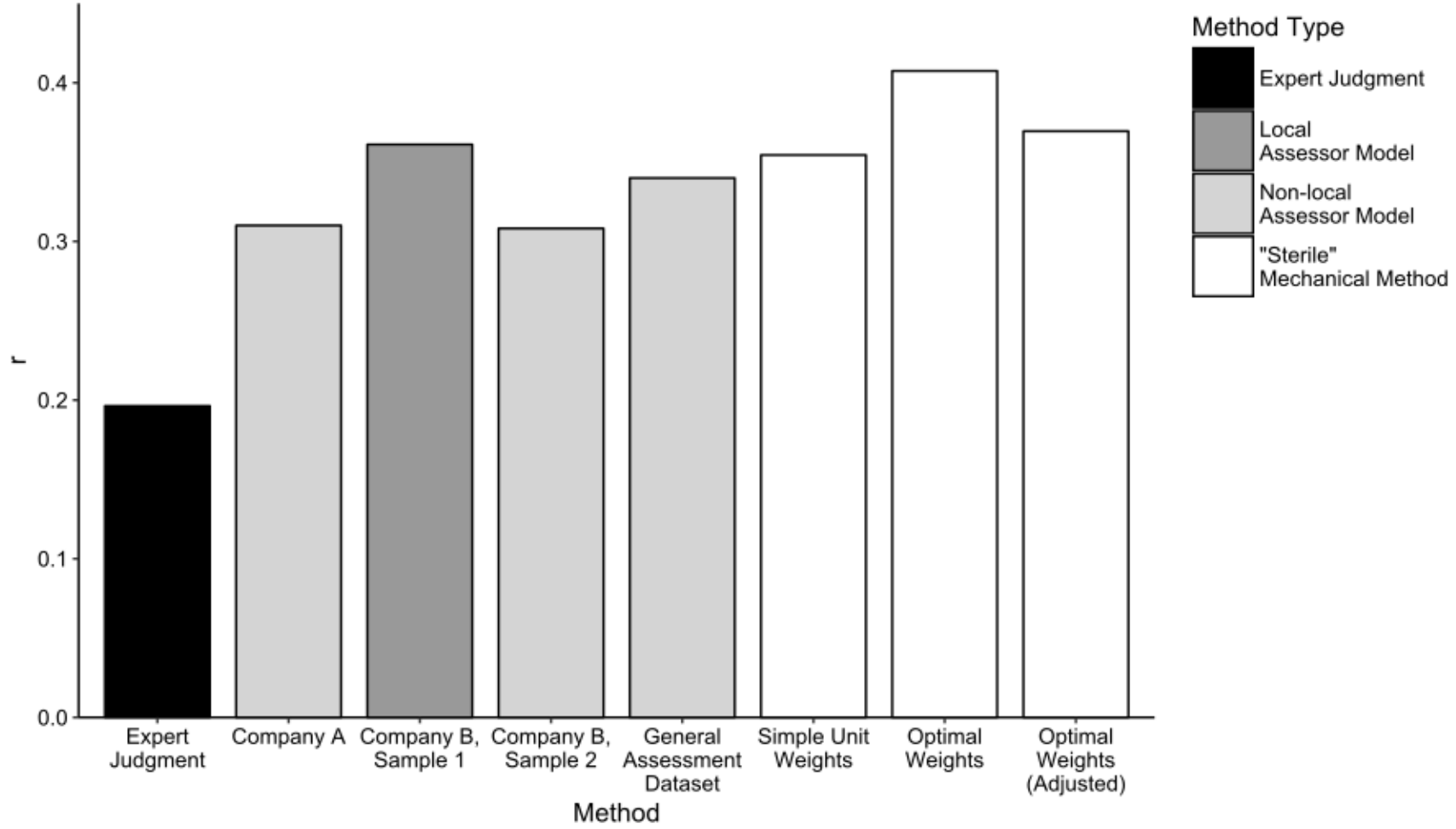

Figure 6. Validities of overall assessment ratings derived from different data combination methods for predicting supervisory ratings of job performance at Company B, Sample 1.

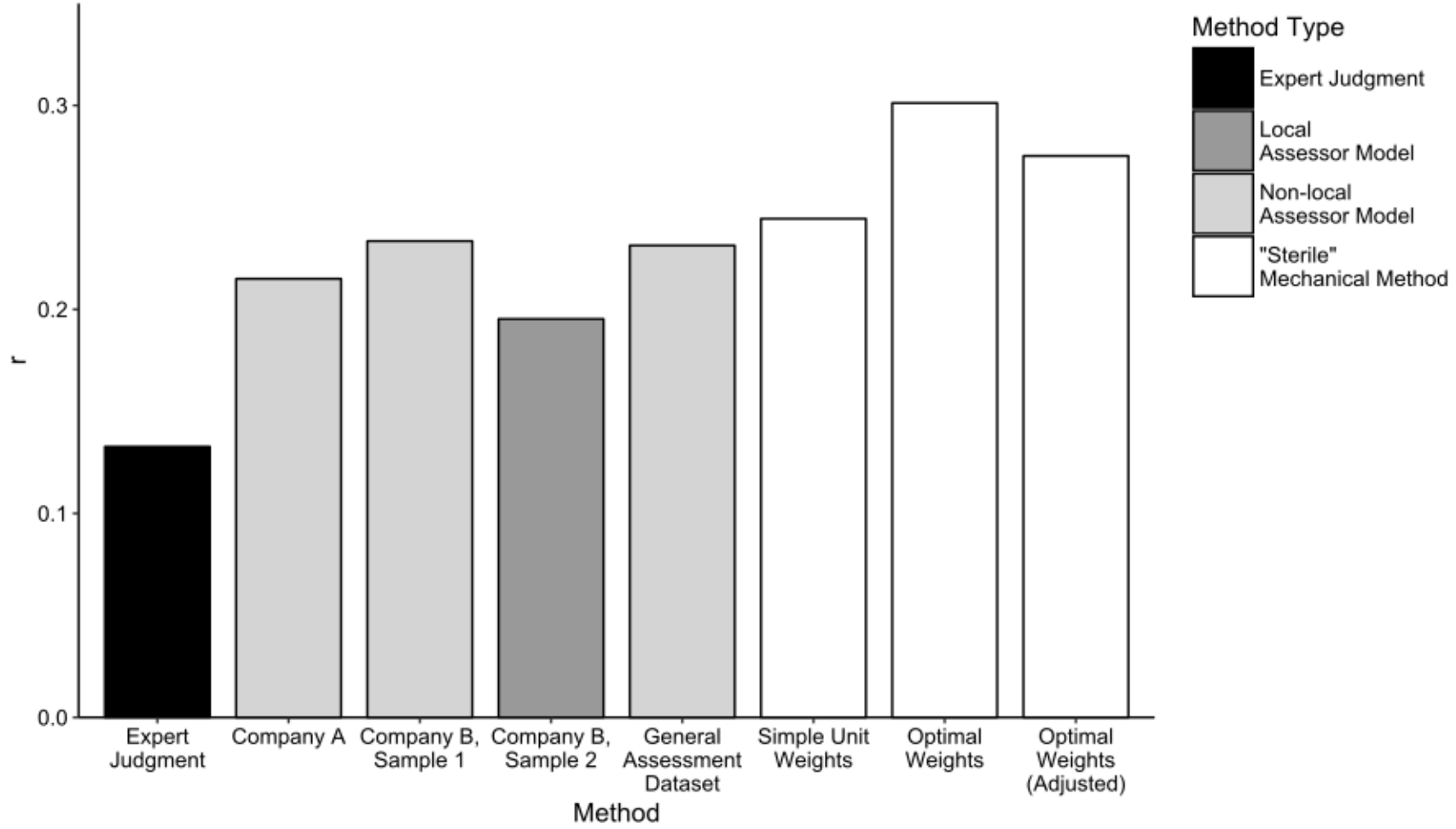

Figure 7. Validities of overall assessment ratings derived from different data combination methods for predicting supervisory ratings of job performance at Company B, Sample 2.

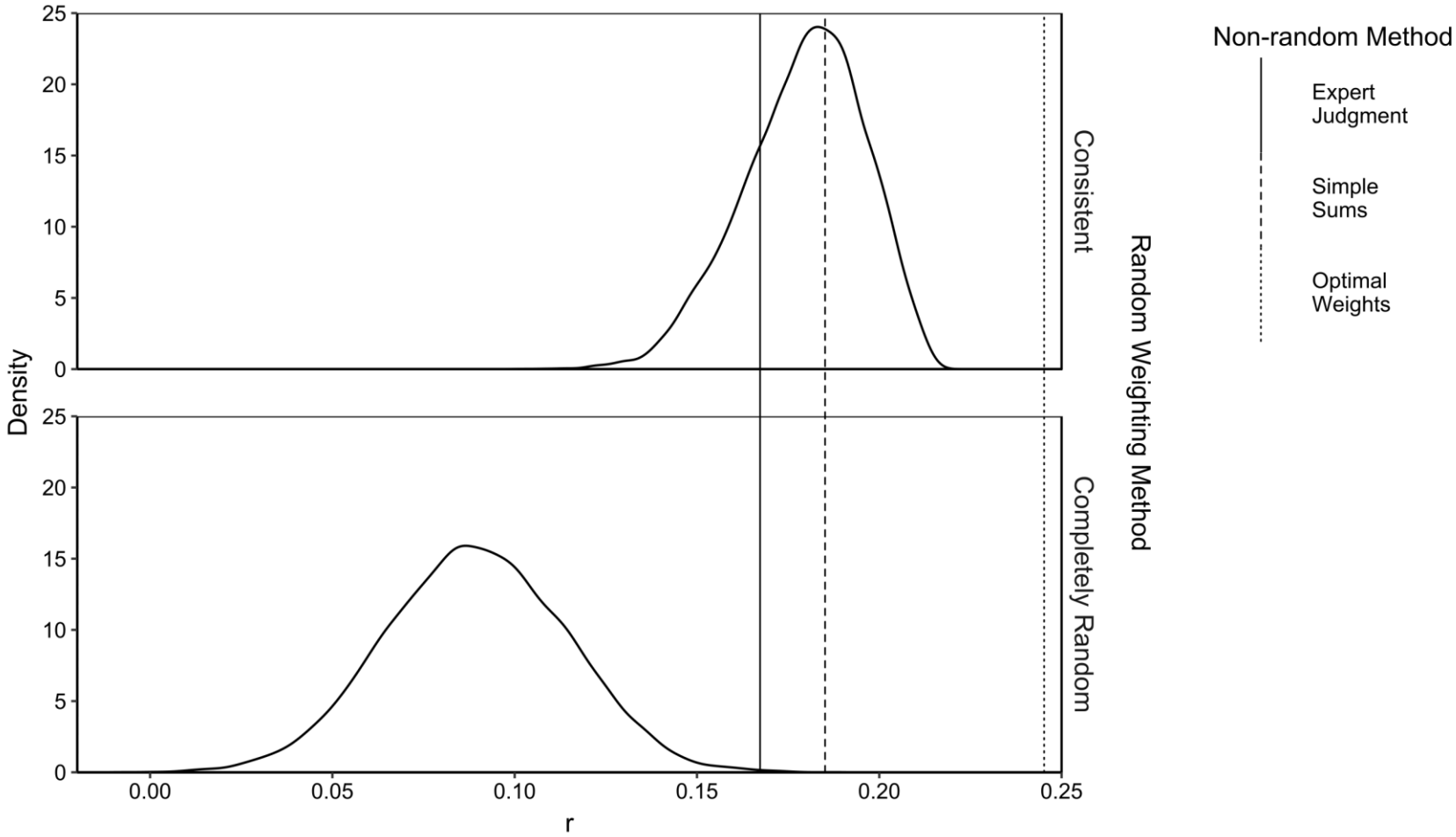

Figure 8. Density distributions of validities (10,000 iterations each) at Company A of predictor scores combined using random positive weights applied consistently across all candidates (top plot) and completely random positive weights (bottom plot) generated for each candidate. Vertical lines are validities at Company A of non-random methods of data combination: expert judgment (solid line), unit weighting via simple sums (dashed line), and optimal weighting (dotted line).

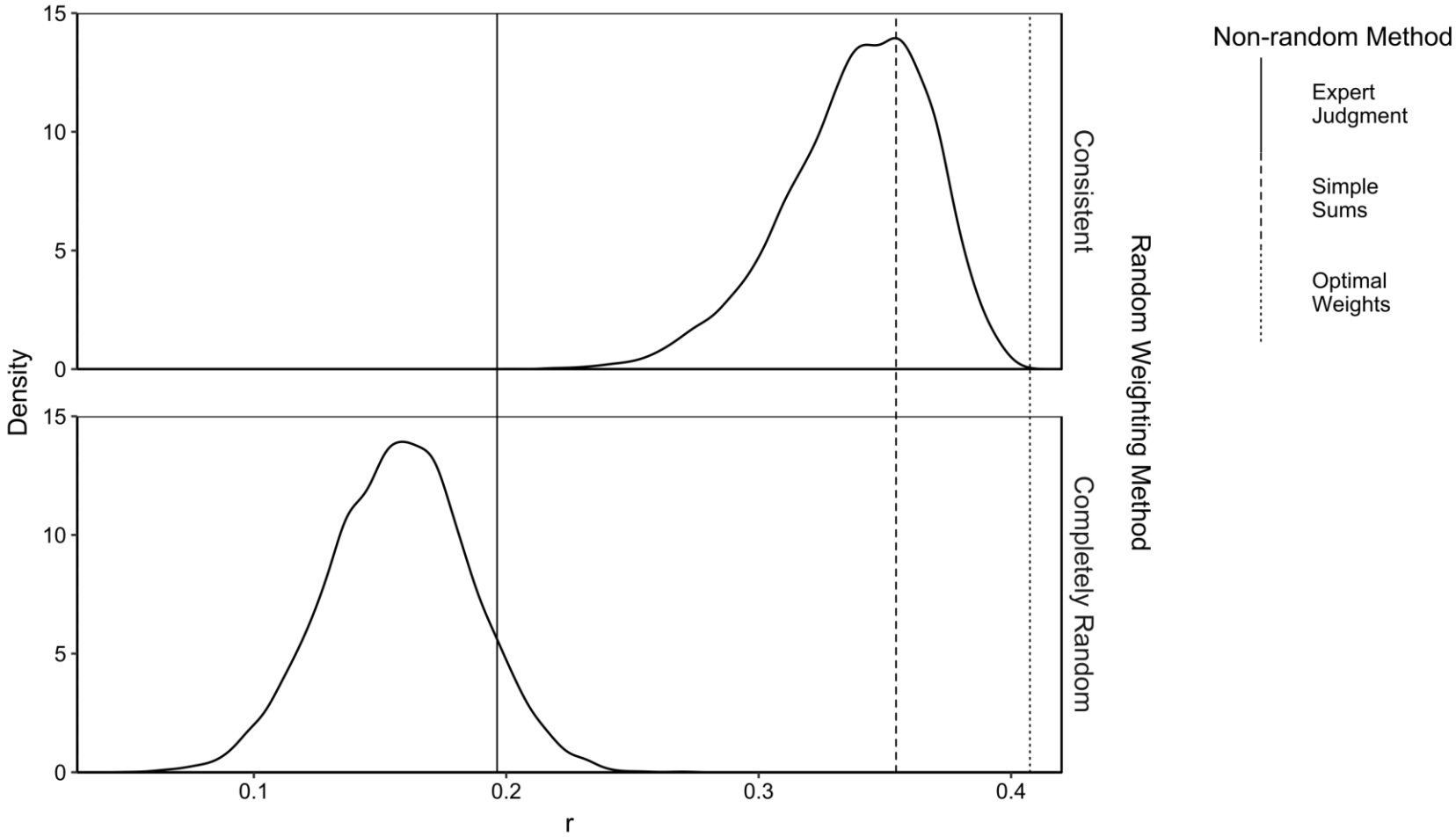

Figure 9. Density distributions of validities (10,000 iterations each) at Company B, Sample 1 of predictor scores combined using random positive weights applied consistently across all candidates (top plot) and completely random positive weights (bottom plot) generated for each candidate. Vertical lines are validities at Company A of non-random methods of data combination: expert judgment (solid line), unit weighting via simple sums (dashed line), and optimal weighting (dotted line).

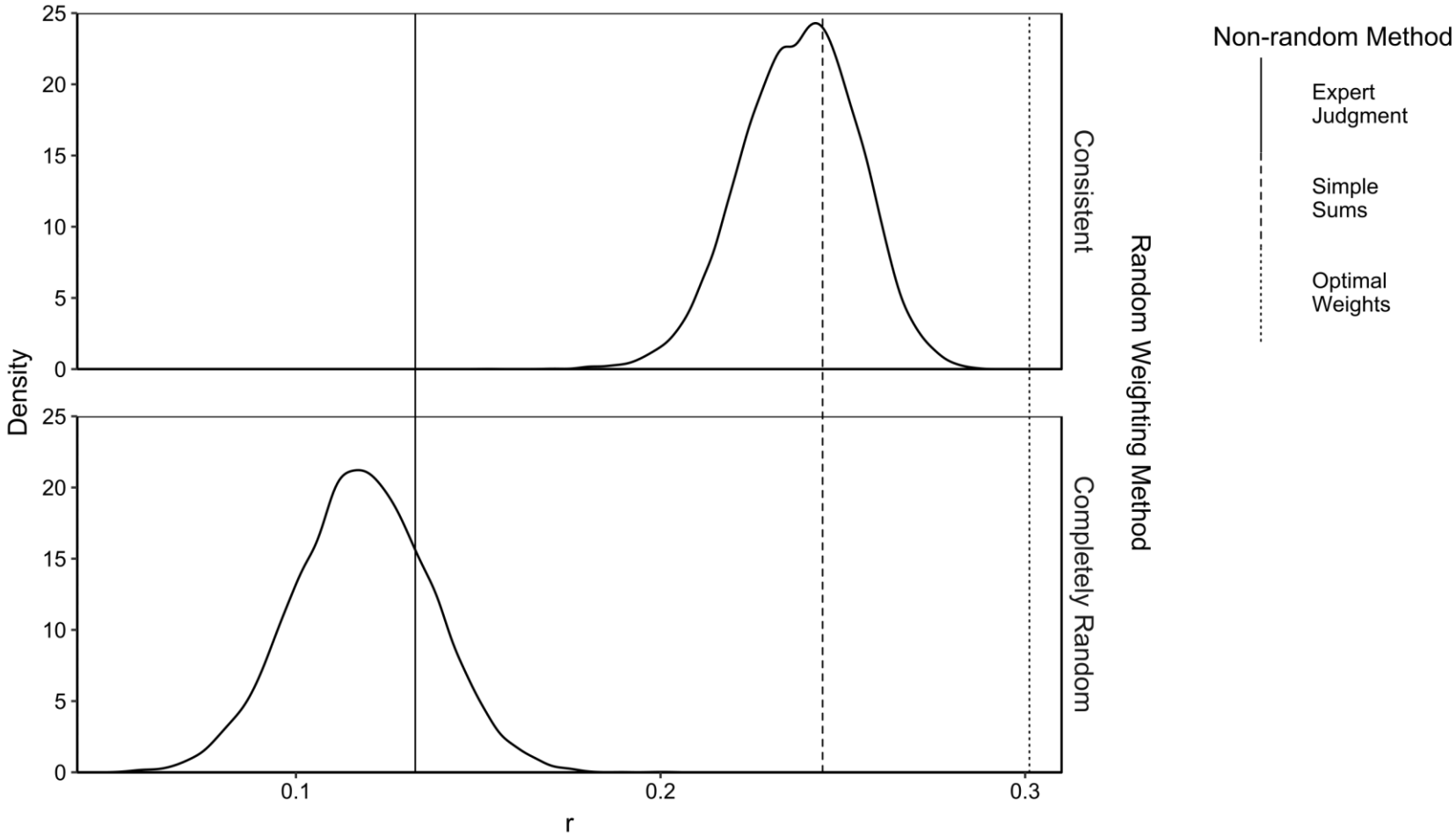

Figure 10. Density distributions of validities (10,000 iterations each) at Company B, Sample 2 of predictor scores combined using random positive weights applied consistently across all candidates (top plot) and completely random positive weights (bottom plot) generated for each candidate. Vertical lines are validities at Company A of non-random methods of data combination: expert judgment (solid line), unit weighting via simple sums (dashed line), and optimal weighting (dotted line).